Progetto Data Technology e Machine Learning

Anno accademico 2017/18

Rima Mirko 793435

Prete Francesco 793389

Spreafico Andrea 793317

INDICE:

DATA TECHNOLOGY

MACHINE LEARNING

1 OBIETTIVI DEL PROGETTO

L'obiettivo principale di questo lavoro è quello di classificare voli aerei a seconda della puntualità del loro arrivo all'aeroporto di destinazione.

Prima di procedere con il lavoro vero e proprio è stato molto importante capire quali fossero le informazioni utili al raggiungimento di tale scopo e, una volta determinate queste informazioni, capire come ottenerle.

Per quanto riguarda le informazioni da ottenere, si è optato per l'utilizzo di tutto ciò che si può conoscere a priori di un volo aereo: aeroporto di partenza, di arrivo, date e orari programmati e una serie di altri valori che saranno analizzati nel dettaglio nelle sezioni successive. Vista l'enorme quantità di voli che hanno luogo ogni giorno, si è inoltre deciso di non utilizzare semplicemente una parte presa a campione di questi voli, ma di selezionare tutti quelli che hanno per aeroporto di origine o destinazione il più grande aeroporto mondiale per traffico aereo: Hartsfield Jackson Atlanta International Airport.

La scelta di fissare una posizione geografica per il nostro studio ci ha indirizzati verso la possibilità di aggiungere le informazioni metereologiche che riguardano questo aeroporto e, più in generale, Atlanta (Georgia, USA).

Data Technology

Raccolta, descrizione e analisi dei dati iniziali

Correzione e aumento della qualità dei dati raccolti

Creazione, descrizione e verifica qualità del dataset finale

Valutazione del modello più appropriato per strutturare il dataset

Machine Learning

Selezione dei dati rilevanti all'interno del dataset generato

Analisi esplorativa del training set scelto

Scelta e descrizione dei modelli di ML utilizzati: DT e NN

Valutazione delle performance dei due modelli


DATA TECHNOLOGY
ATN_STAGE_FK (FK)
ATN_STAGE_MODIFIER_FK (FK)
ATN_PUBLIC_ID
ANA_NODE
ANO_OID (FK)
ANO_SPECIES_FK (FK)
ANO_COMPONENT_NAME
ANO_IS_PRIMARY
ANO_IS_GROUP
ANO_PUBLIC_ID
ANO_DESCRIPTION
anode_obj
asyn_obj
ANA_OBJECT
OBJ_OID
OBJ_CREATION_DATETIME
OBJ_CREATOR_FK
arel_obj
asrc_obj
stg_obj
aeditor_obj
attrib_obj
ANA_RELATIONSHIP
REL_OID (FK)
REL_RELATIONSH
REL_CHILD_FK (F
REL_PARENT_FK
SEQUENCE
REF_SPECIES
RSP_NAME
RSP_LATIN_NAME
PREFIX
ANA_STAGE
STG_OID (FK)
STG_SPECIES_FK (FK)
STG_NAME
STG_SEQUENCE
STG_DESCRIPTION
STG_SHORT_EXTRA_TEXT
STG_PUBLIC_ID
ANA_ATTRIBUTION
ATR_OID (FK)
ATR_OBJECT_FK (FK)
ATR_SOURCE_FK (FK)
ATR_EVIDENCE_FK (FK)
ATR_COMMENTS
ANA_SOURCE
SRC_OID (FK)
SRC_NAME
SRC_AUTHORS
SRC_FORMAT_FK (FK)
SRC_YEAR

2 DATA UNDERSTANDING

2.1 Raccolta dati iniziali

Inizialmente si è cercato di trovare le informazioni necessarie al lavoro in dataset pronti per l'utilizzo ma, purtroppo, con scarsi risultati. Allora il lavoro è stato separato in due obiettivi: da un lato trovare le informazioni relative alle condizioni metereologiche di Atlanta, dall'altro trovare le informazioni riguardo i voli che transitano per l'aeroporto da noi identificato come caso di studio.

In questo modo siamo riusciti ad ottenere due dataset contenenti, tra le altre, le informazioni a noi necessarie. In particolare:

- ***Weather_ATL*:** un database contenente informazioni relative alle condizioni meteo nella città di Atlanta (in particolare nel distretto dell'aeroporto) nell'anno solare 2015.
  (data source: http://www.iweathernet.com/atlanta-weather-records)

- ***Flights_delay*:** un database riguardante tutti i voli in arrivo e in partenza dai maggiori aeroporti americani nell'anno solare 2015.
  (data source: https://www.kaggle.com/usdot/flight-delays)

Questi due dataset rappresentano i dati raw dal quale parte la fase di data mining finalizzata ad ottenere un unico dataset con le sole informazioni utili al nostro scopo.

2.2 Descrizione dei dati iniziali

In questa sezione saranno analizzati nel dettaglio i due dataset e tutte le informazioni in essi contenute.

2.2.1 Weather_ATL

Il database contiene 365 istanze, una per ogni giorno dell'anno, in cui sono rappresentate attraverso 10 attributi le condizioni metereologiche della giornata. Nella tabella sottostante è riportata la descrizione degli attributi del dataset.

Attributo	Tipo	Descrizione
Date	Date	Data in considerazione nel formato MM/GG/AAAA.
MaximumTemp	Int	Temperatura massima rilevata (° F).
MinimumTemp	Int	Temperatura minima rilevata (° F).
AverageTemp	Double	Media delle rilevazioni giornaliere (° F).
DepartureTemp	Double	Differenza tra la temperatura media giornaliera e il valore normale del periodo (° F).
HDD	Int	Il numero di gradi in cui la temperatura media giornaliera è inferiore a 65° F.

CDD	Int	Il numero di gradi in cui la temperatura media giornaliera è superiore a 65° F.
Precipitation	Char	Precipitazioni medie, in inches all'ora.
NewSnow	Double	Inches di neve caduta durante il giorno considerato.
SnowDepth	Int	Media degli inches di neve al suolo.

2.2.2 Flights_delay

Il database contiene 5819079 istanze riguardanti tutti i voli aerei avvenuti nei maggiori aeroporti americani. Ognuna di queste tratte è descritta dai 31 attributi riportati e descritti nella tabella sottostante.

Attributo	Tipo	Descrizione	Range
Year	Int	Anno in considerazione.	2015
Month	Int	Mese in considerazione.	1-12
Day	Int	Giorno in considerazione.	1-7
Day of Week	Int	Giorno della settimana preso in considerazione. 1 indica "lunedì", 2 "martedì" ecc.	1-7
Airline	Char	IATA Code della compagnia aerea.	/
Flight_number	Int	Numero identificativo del volo.	/
Tail_number	Char	Numero identificativo del velivolo.	/
Origin_Airport	Char	IATA Code dell'aeroporto di partenza.	/
Destination_Airport	Char	IATA Code dell'aeroporto di arrivo.	/
Scheduled_Departure	Time*	Orario di partenza pianificato.	0 - 2359
Departure_Time	Time	Orario effettivo di partenza. È calcolato come: wheels_off - taxi_out	0 - 2359
Departure_Delay	Int	Minuti di ritardo o anticipo della partenza. È calcolato come: departure_time - scheduled_departure	/
Taxi_Out	Int	Minuti trascorsi dalla chiusura del gate al distacco delle ruote dalla pista.	/
Wheels_Off	Time	Orario in cui le ruote del velivolo si staccano dal suolo durante la partenza.	0 - 2359

Scheduled_Time	Int	Tempo teorico trascorso dalla chiusura del gate dell'aeroporto di partenza fino all'arrivo al gate dell'aeroporto di destinazione.	/
Elapsed_Time	Int	Tempo reale trascorso dalla chiusura del gate dell'aeroporto di partenza fino all'arrivo al gate dell'aeroporto di destinazione. È calcolato come: air_time + taxi_in + taxi_out	/
Air_Time	Int	Tempo effettivo di volo. È calcolato come: wheels_on - wheels_off	/
Distance	Int	Distanza (in miglia) tra l'aeroporto di partenza e quello di destinazione.	/
Wheels_On	Time	Orario in cui le ruote del velivolo toccano il suolo durante l'atterraggio.	0 - 2359
Taxi_In	Int	Minuti trascorsi dal momento in cui le ruote toccano il suolo fino all'arrivo al gate dell'aeroporto di destinazione.	/
Scheduled_Arrival	Time	Orario di arrivo pianificato.	0 - 2359
Arrival_Time	Time	Orario effettivo di arrivo. È calcolato come: wheels_on + taxi_in	0 - 2359
Arrival_Delay	Int	Minuti di ritardo o anticipo all'arrivo. È calcolato come: arrival_time - scheduled_arrival	/
Diverted	Boolean	Flag che indica se il volo è stato, per qualche motivo, dirottato. Il valore 1 indica un volo dirottato.	0 - 1
Cancelled	Boolean	Flag che indica se il volo è stato, per qualche motivo, cancellato. Il valore 1 indica un volo cancellato.	0 - 1
Cancellation_Reason	Char	Carattere che indica il motivo di un'eventuale cancellazione del volo. I possibili valori sono: (A) Air Carrier; (B) Extreme Weather; (C) National Aviation System (NAS); e (D) Security.	A - D + NAS
Air_System_Delay	Int	Minuti di ritardo causati dal sistema di areazione del velivolo.	/
Security_Delay	Int	Minuti di ritardo causati da motivi di sicurezza.	/
Airline_Delay	Int	Minuti di ritardo causati dalla compagnia aerea.	/
Late_Aircraft_Delay	Int	Minuti di ritardo causati da altri velivoli in ritardo.	/
Weather_Delay	Int	Minuti di ritardo causati dal maltempo.	/

* il tipo di dato "Time" è nel formato xxyy con xx ora (da 00 a 24) e yy minuti (da 00 a 59).

2.3 Selezione dei dati rilevanti

Dopo aver individuato i dataset da utilizzare si è reputato necessario, prima di ogni altra cosa, filtrare i dati a causa della massiccia quantità di istanze e attributi non rilevanti ai nostri scopi.

Per quanto riguarda il dataset *Weather_ATL*, esso conteneva tutte e sole le istanze necessarie allo studio, ma con attributi che o non erano significativi (come quelli relativi alla neve, considerando che in tutto l'anno solare 2015 non si sono verificate nevicate ad Atlanta) o non aggiungevano informazioni utili (come *"HDD"* e *"CDD"*, che sono semplici valori che indicano il discostarsi della media giornaliera da un valore fissato). Abbiamo quindi deciso di selezionare gli attributi considerati rilevanti, ottenendo in tal modo un sottoinsieme dei dati rinominato *Weather_ATL_Data_Quality* con la totalità delle istanze di *Weather_ATL,* ma solamente alcune colonne. In particolare:

"Date", "MaximumTemp", "MinimumTemp" e "Precipitation".

Come suggerisce il nome, questo dataset è quello utilizzato nella fase di verifica della qualità dei dati che sarà trattata nell'apposito capitolo (2.5)

La situazione era molto più complessa nel dataset *Flights_delay*. In questo caso non solo molti attributi erano poco significativi, ma alcuni erano registrati durante e dopo il volo (è il caso di *"Taxi_In", "Taxi_Out"*... ecc.). Come scelto inizialmente si è deciso di considerare solamente le informazioni che si potevano conoscere prima della partenza dell'aereo e, quindi, anche in questo caso si è costruito un nuovo dataset considerando una parte degli attributi. Quelli selezionati risultano essere:

"Month", "Day", "Day_of_Week", "Airline", " Origin_Airport", "Destination_Airport", "Scheduled_Departure", "Scheduled_Time", "Distance", "Scheduled_Arrival",

Inoltre si è reputato necessario l'utilizzo dei seguenti attributi, il cui valore non risulta disponibile prima della partenza, con lo scopo di condurre un'accurata fase di data quality.

"Arrival_Time" e "Arrival_Delay".

In questo caso, però, la fase di selezione non si limita alle colonne: come indicato nella fase di descrizione dei dataset, in *Flights_delay* sono contenuti tutti i voli dei maggiori aeroporti per un totale di quasi 6 milioni di istanze. Prima di salvare il nuovo database sono state filtrate tutte le righe che non comprendevano né nell'aeroporto di origine, né in quello di destinazione, il codice identificativo dell'aeroporto di Atlanta (codice IATA -> "ATL").

L'output di questa fase risulta essere, anche per quanto riguarda i voli, un nuovo dataset su cui effettuare lo studio riguardante la qualità dei dati: *Flights_delay_ATL_Data_Quality*.

2.4 Esplorazione dei dati

Prima di procedere con la fase di valutazione dei due dataset appena creati, facciamo un'analisi inziale in modo da comprendere meglio i dati e scoprire eventuali correlazioni tra i valori di questi ultimi. Questa analisi, divisa secondo il dataset utilizzato, è riportata di seguito.

Tutte le statistiche sono state effettuate utilizzando R ed è possibile visionare il file in "Data_Technology\Code\2_Data Exploration\Explore_Data.R"

2.4.1 Dataset *Flights_delay_ATL_Data_Quality*

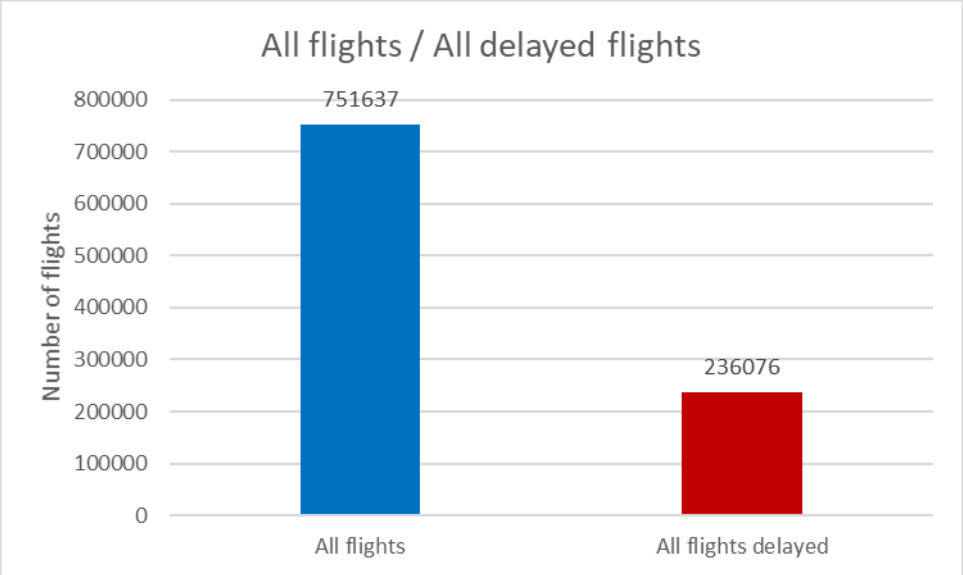
Inizialmente il dataset è stato analizzato per quanto riguarda la distribuzione dei voli nei diversi giorni della settimana e la distanza percorsa in ogni tratta. Durante l'esplorazione dei dati sono state individuate 7285 righe con valori NA, corrispondenti ai voli cancellati o dirottati. Nelle tabelle, a scopo informativo, sono stati indicati i valori NA divisi nelle diverse categorie, tuttavia i valori considerati per le statistiche sono al netto di questi valori.

All flights / All delayed flights

Analizzando la distribuzione dei voli nel dataset, abbiamo cercato il numero totale dei voli e il numero totale dei voli in ritardo, con lo scopo di comprendere il rapporto tra i due valori.

I risultati sono mostrati nella tabella e nel grafico qui riportati.

All flights	All flights delayed	NA
751637	236076	7285

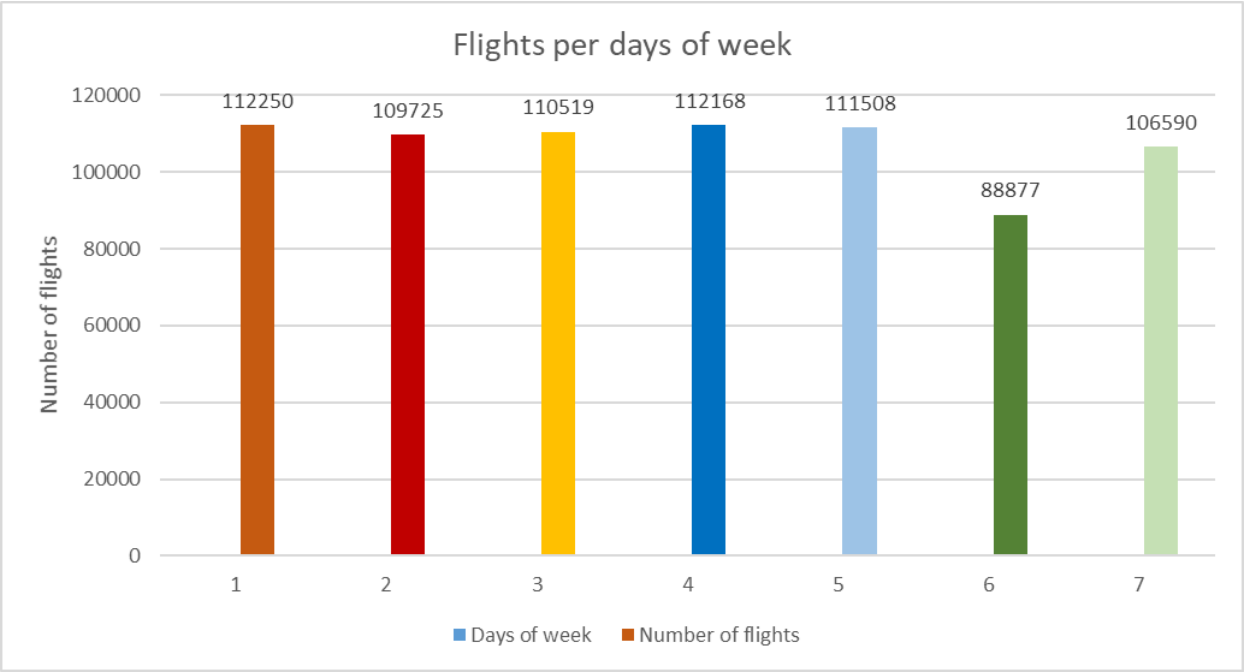


Dopo aver ottenuto i dati abbiamo concluso che la percentuale dei voli in ritardo sul totale dei voli è circa del 31%.

Days of Week

Analizzando la distribuzione dei voli nei vari giorni della settimana, abbiamo cercato eventuali anomalie nell'uniformità della distribuzione. Per questo motivo abbiamo raggruppato i voli per giorno della settimana ottenendo i risultati sottostanti.

Days of week	Number of flights	NA
1	112250	1186
2	109725	1507
3	110519	1444
4	112168	1238
5	111508	629
6	88877	634
7	106590	647

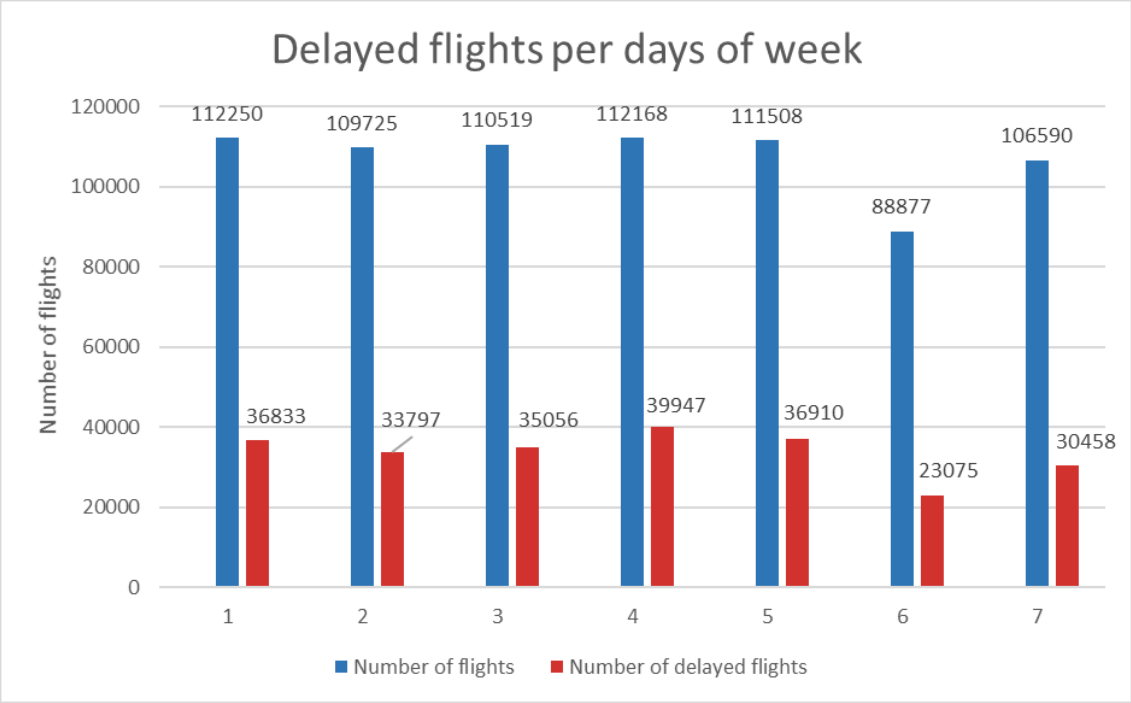


Days of week enum						
1	2	3	4	5	6	7
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday

Si può notare come il numero di voli sia bilanciato nei giorni della settimana ad eccezione del sabato dove sono presenti meno voli (circa il 20% in meno sulla media degli altri giorni della settimana).

A questo punto si è cercata una correlazione tra i ritardi e il giorno della settimana in cui è partito il volo, ottenendo i dati qui riportati.

Days of week	Number of flights	Number of delayed flights	NA
1	112250	36833	1186
2	109725	33797	1507
3	110519	35056	1444
4	112168	39947	1238
5	111508	36910	629
6	88877	23075	634
7	106590	30458	647



Osservando i dati si è concluso che la proporzione tra i voli divisi per giorni e i ritardi è mostrata nella tabella sottostante

Percentage of delayed flights per day						
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
33%	31%	32%	36%	33%	26%	29%

Si nota che la percentuale tra i ritardi dei vari giorni si aggira sul 31%, con pochi giorni (sabato picco più basso e giovedì picco più alto) che si allontanano di poco da questa media che tuttavia resta abbastanza bilanciata nei vari giorni.

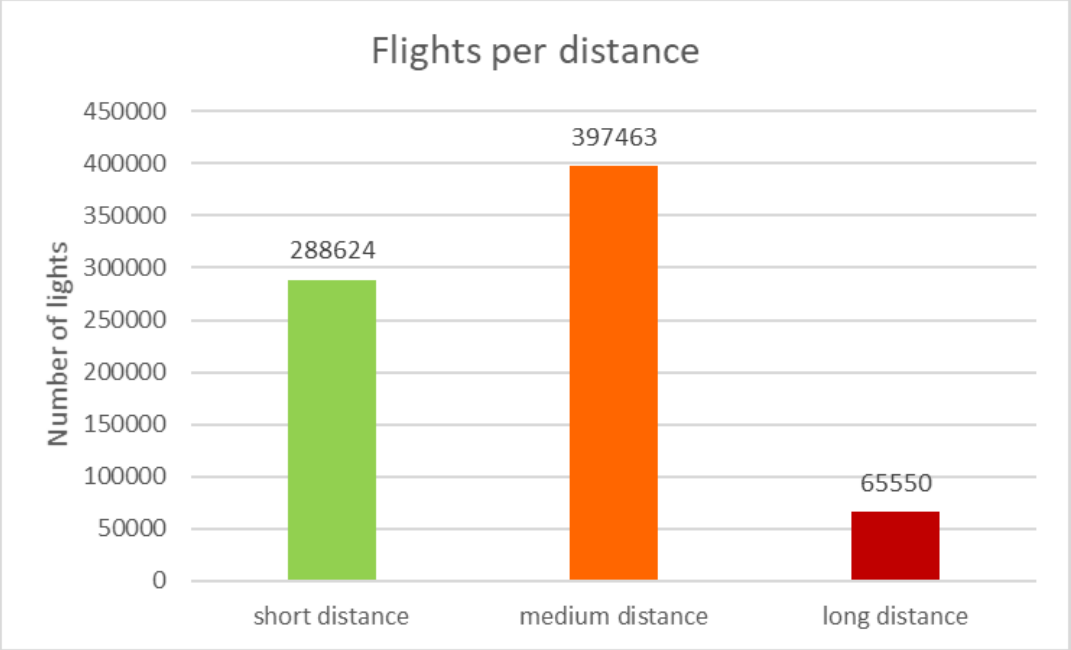
Flights per distance

Un altro fattore importanti ai fini della comprensione del dataset riguarda la lunghezza dei voli. Considerando la lunghezza da costa a costa degli USA e che i voli sono tutti nazionali, abbiamo deciso di suddividerli in tre categorie:

range of distances (miles)		
short distance	medium distance	long distance
≤ 450	450 < x ≤ 1500	> 1500

I risultati di queste analisi sono riportati con i relativi grafici qui di seguito.

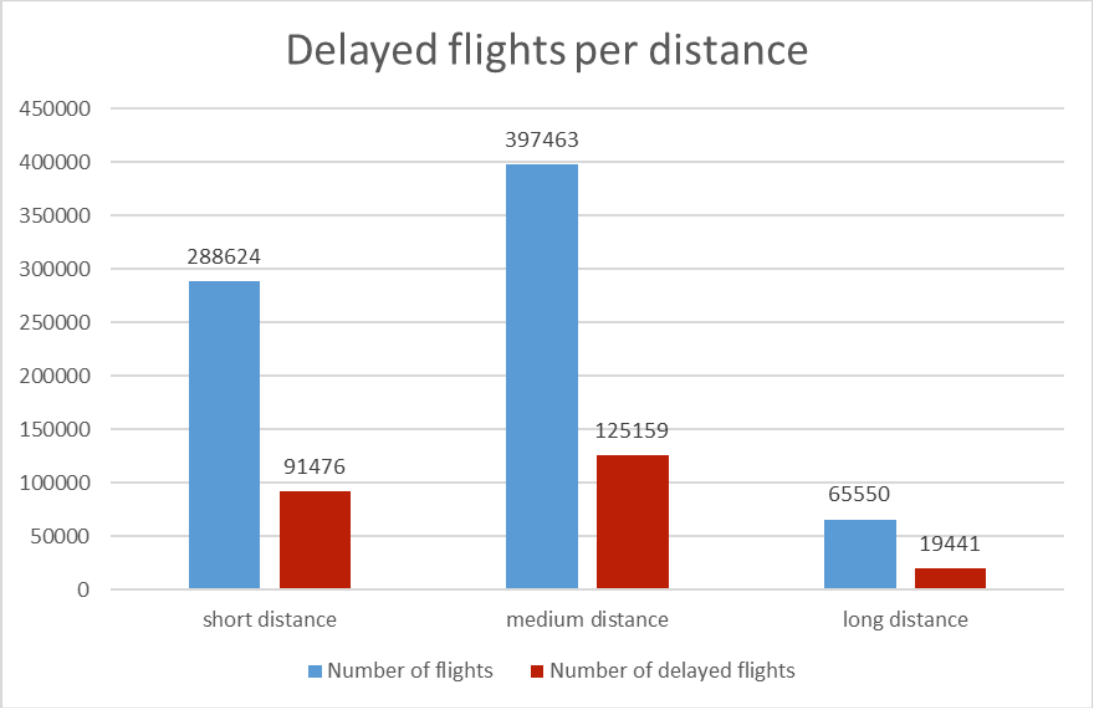
Type of distance	Number of flights	NA
short distance	288624	2205
medium distance	397463	4696
long distance	65550	384



I voli di breve distanza corrispondono al 38% dei voli totali, quelli di media e lunga distanza, rispettivamente al 53% e al 9%.

Anche in questo caso si è poi provato a considerare i ritardi sulle tratte per provare a trovare una correlazione tra la distanza e i ritardi.

Type of distance	Number of flights	Number of delayed flights	NA
short distance	288624	91476	2205
medium distance	397463	125159	4696
long distance	65550	19441	384



Ottenuti i dati si è notato di come i ritardi siano distribuiti quasi uniformemente sulle diverse distanze, restando vicini alla media dei ritardi totali del 31%.

Percentage of delayed flight per distance		
short distance	medium distance	long distance
32%	31%	30%

2.4.2 Dataset *Weather_ATL*

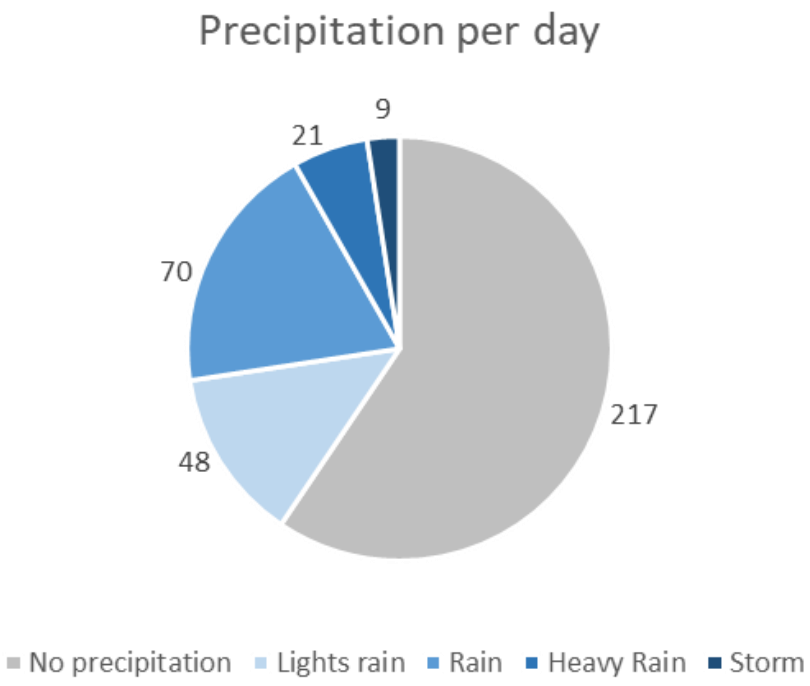
Per quanto riguarda le condizioni metereologiche sono state effettuate le due statistiche riguardano i dati relativi alle temperature e alle precipitazioni.
È stato deciso di ignorare la neve come fattore climatico utile al nostro lavoro poiché ad Atlanta non sono state registrate nevicate nel 2015.

Precipitation per day
In questa fase abbiamo deciso di dividere i valori relativi alle precipitazioni in 5 gruppi in base al livello medio di pollici all'ora di pioggia caduti nella giornata:

Inches per hour				
No precipitation	Light rain	Rain	Heavy Rain	Storm
0	$0 < x \leq 0{,}1$	$0{,}1 < x \leq 0{,}79$	$0{,}79 < x \leq 1{,}5$	$x > 1{,}5$

I risultati ottenuti sono riportati qui di seguito.

Precipitation type	Number of days
No precipitation	217
Light rain	48
Rain	70
Heavy rain	21
Storm	9



Come si può evincere dai dati ottenuti, nella maggior parte dell'anno non sono state rilevate precipitazioni. Inoltre nei giorni piovosi raramente le precipitazioni si sono rilevate intense.

Percentage of the dataset division				
No precipitation	Light rain	Rain	Heavy rain	Storm
59,5%	13%	19%	6%	2,5%

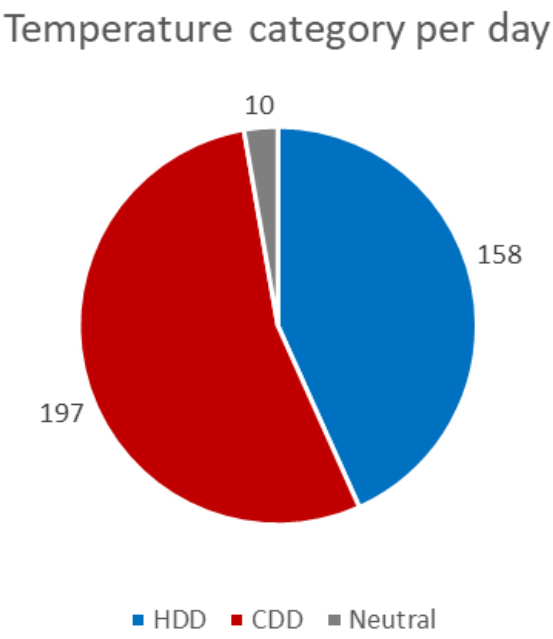
Temperature

Successivamente abbiamo spostato l'attenzione sulle temperature considerando i valori *CDD* e *HDD* ovvero i valori in cui la temperatura media della giornata si discosta dal valore di 65° Fahrenheit (circa 18.3° C).
In questo modo abbiamo identificato 3 categorie:

Types of temperature		
HDD	CDD	Neutral
< 65 F°	> 65 F°	= 65 F°

Ancora una volta sono riportati di seguito i grafici con i risultati ottenuti.

Category	Number of days
HDD	158
CDD	197
Neutral	10



Il dataset è stato quindi diviso in HDD che comprende il 43% dei nostri dati sulle temperature, CDD e Neutral rispettivamente 54% e 3%.

2.5 Verifica della qualità dei dati

Terminata la fase esplorativa dei dati siamo passati alla verifica della qualità dei due dataset filtrati. Prima di procedere con questa analisi, però, è stato necessario determinare quali, tra le possibili dimensioni di qualità, utilizzare nell'analisi.

La scelta è ricaduta su Completezza, Accuratezza, Consistenza e Unicità. Inoltre, per ognuna di queste dimensioni è stata decisa una metrica da utilizzare nella misurazione, facendo riferimento *a "Methodologies for Data Quality Assessment and Improvement"* (Carlo Batini at all, 2009). La tabella sottostante riporta le dimensioni e le relative metriche utilizzate.

Dimensione	Tipo	Metrica
Completezza	Completezza della tabella	Numero di valori non nulli / Numero totale di valori
Accuratezza*	Accuratezza sintattica	Numero di valori corretti / Numero totale di valori
Consistenza	Integrità referenziale	Numero di valori consistenti / Numero totale di valori
	Consistenza dei dati**	Numero di valori consistenti / Numero totale di valori
Unicità	Unicità delle tuple	Numero di tuple duplicate

* Per quanto riguarda l'accuratezza sintattica, con valore corretto si intende un valore che, cercato in un file contenente tutti e soli i valori validi, viene trovati.

** Per consistenza dei dati si intende un controllo sulla correttezza dei valori numerici ricavabili attraverso una formula. Nel caso specifico l'unico attributo di questo tipo risulta essere "Arrival_Delay" = "Arrival_Time" – "Scheduled_Arrival".

Nelle prossime sezioni saranno riportati i risultati ottenuti nei dataset su cui è stata effettuata la verifica della qualità dei dati.

2.5.1 Dataset *Flights_delay_ATL_Data_Quality*

Dimensione	Risultati
Completezza della tabella	9094287 / 9107064 = 0.999
Accuratezza sintattica	*"Airline"*: 758922 / 758922 = 1 *"Destination"*: 693740 / 758922 = 0.914 *"Origin"*: 693740 / 758922 = 0.914 Mean = 0.943
Integrità referenziale	*"Airline"* 758922 / 758922 = 1 *"Scheduled_Arrival"* 758922 / 758922 = 1 *"Day"* 758922 / 758922 = 1 *"Month"* 758922 / 758922 = 1 *"Day_of_week"* 758922 / 758922 = 1 *"Arrival_Delay"* 758922 / 758922 = 1 *"Scheduled_Departure"* 758922 / 758922 = 1 *"Distance"* 758922 / 758922 = 1 *"Scheduled_Time"* 758922 / 758922 = 1 *"Destination_Airport"* 693740 / 758922 = 0.914 *"Origin_Airport"* 693740 / 758922 = 0.914 Mean = 0.984
Consistenza dei dati	751.637 / 758922 = 0,990
Unicità delle tuple	758922 - 758921 = 1 tupla duplicata

Nb. è possibile prendere visione del codice R che genera i seguenti risultati nel file "Data_Technology\Code\3_Data Quality\Data_Quality_Flights_delay.R"

Come spiegato nella fase "Selezione dei dati rilevanti" il dataset utilizzato in questa fase è un sottoinsieme di quello originale. Da qui, deriva il primo errore riscontrato: l'eliminazione di alcune colonne non ritenute utili ai fini del lavoro ha eliminato il codice identificativo univoco del volo (*"Flights_Number"*) andando a eliminare la discriminazione tra due delle righe del dataset, creando la tupla duplicata.

In realtà il dataset *Flights_delay_ATL_Data_Quality* non è composto solo dalle istanze con codice IATA uguale a "ATL", questo perché dalla fase di verifica dell'integrità referenziale è stato riscontrato che non tutte le tuple riportavano come valore in *"Destination_Airport"* e *"Origin_Airport"* un codice IATA a 3 cifre: circa il 9% dei valori era composto a un codice a 5 cifre che non rispecchiava nessun valore IATA valido. Dopo alcune ricerche è stata scoperta una corrispondenza univoca tra codici IATA e i codici numerici a 5 cifre (visibile nel file "Data_Technology\Datasets\Airports_codes_translations.csv"). Uno tra questi nuovi

codici, inoltre, risultava corrispondere con l'aeroporto di Atlanta considerato in questo studio e si è quindi deciso di inserire nel dataset per la valutazione della qualità dei dati anche queste nuove istanze, senza però modificarne il valore dal codice numerico al codice IATA corrispondente. Sono quindi questi nuovi valori inseriti, ma non modificati, a spiegare l'errore nell'integrità referenziale.
Il passo successivo è stato controllare la validità dei valori degli attributi non numerici del dataset: *"Destination_Airport", "Origin_Airport" e "Airline".* Per fare ciò ci siamo serviti di una lista di codici IATA validi per aeroporti e compagnie aeree e abbiamo controllato che i valori del dataset rispecchiassero uno dei valori validi (nei file "Data_Technology\Datasets\Airports.csv" e "Data_Technology\Datasets\Airlines.csv"). Come si può vedere dai risultati i codici sbagliati sono in quantità uguale a quelli errati nella dimensione dell'integrità referenziale: questo indica che, tutti i codici a 5 cifre sono errati in entrambe le dimensioni ma anche che, i codici IATA a 3 cifre presenti nel dataset erano tutti codici validi.

Un'ultima osservazione riguarda i valori "null": la percentuale di valori non nulli risulta identica a quella dei valori consistenti. Questo è dovuto al fatto che per i voli dirottati e / o cancellati, non è riportato il valore dell'attributo *"Arrival_Delay"*, in quanto un volo che non è mai arrivato a destinazione non può avere un ritardo di arrivo. Questo fattore influisce negativamente anche la consistenza dei dati, rendendo impossibile verificare il corretto calcolo del ritardo ("*Arrival_Delay" = "Arrival_Time" – "Scheduled_Arrival")* e quindi della consistenza.

2.5.2 Dataset *Weather_ATL_Data_Quality*

Dimensione	Risultati
Completezza della tabella	365 / 365 = 1
Integrità referenziale	365 / 365 = 1
Consistenza dei dati	365 / 365 = 1
Unicità delle tuple	365 - 365 = 0 tuple duplicate

Nb. è possibile prendere visione del codice R che genera i seguenti risultati nel file "Data_Technology\Code\3_Data Quality\Data_Quality_Weather.R"

Dai dati riportati per questo dataset è possibile notare una particolarità: la mancanza dei risultati che riguardano l'accuratezza sintattica. Questo è dovuto al fatto che i valori del dataset sono interamente numerici e un controllo sintattico su tali valori non porterebbe ad alcun risultato.

Di tali valori numerici, però, è stato effettuato un controllo sulla consistenza dei dati tra loro correlati: la temperatura minima giornaliera deve essere effettivamente inferiore a quella massima dello stesso giorno. Questo controllo ha avuto un risultato positivo per tutte le istanze del dataset.

3 DATA CLEANING

3.1 Correzione dei dati

Terminata la fase di analisi della qualità dei dati, si sono riparate tutte le lacune qualitative attraverso del codice R visualizzabile nei file relativi alla creazione dei dataset ("Data_Technology\Code\1_Data Generation\ Flights_Datasets_Elabs.R" e "Data_Technology\Code\1_Data Generation\ Weather_Datasets_Elabs.R").

Il dataset *Weather_ATL_Data_Quality* è rimasto pressoché invariato, a meno di una modifica utile per la fase successiva del progetto. Essa consiste nell'eliminazione dell'attributo *"Date"* in favore di *"Month"* e *"Day"*, in modo da rendere uniforme, nei due dataset, il formato della data. Non è stato considerato l'attributo *"Year"* poiché costante tra tutte le istanze. Effettuato questo passaggio, si è giunti facilmente alla versione definitiva di questo dataset: *Weather_ATL_Restored.*

Ancora una volta, la parte complessa e onerosa, riguarda il file *Flights_delay_ATL_Data_Quality*. Saranno ora trattate, una dimensione di qualità alla volta, le strategie risolutive per ogni problema riscontrato nel dataset.

- I valori "null" sono stati rimossi riducendo la completezza di rappresentazione degli oggetti del mondo reale (i voli aerei) dal 100% al 99%.
- Gli errori di accuratezza sintattica, come già considerato precedentemente sono strettamente legati a quelli di integrità referenziale: applicando la sostituzione di tutti i codici a 5 cifre con i relativi codici IATA, entrambi gli errori saranno corretti (tutti i valori torneranno a rispettare l'integrità referenziale e di conseguenza saranno anche contenuti nella lista dei valori validi possibili per il codice IATA di un aeroporto).
- Anche per quanto riguarda la consistenza dei dati, l'errore è in automatico riparato con l'eliminazione dei valori "null": tutti e soli i dati con *"Arrival_Delay"* nullo risultavano errati. Rimuovendo queste tuple i restanti valori, sono coerenti.

Terminate le modifiche, un nuovo sottoinsieme dei dati di *Flights_delay_ATL_Data_Quality è stato salvato con il nome di Flights_delay_ATL_Restored*, andando a costituire il dataset finale dei voli.

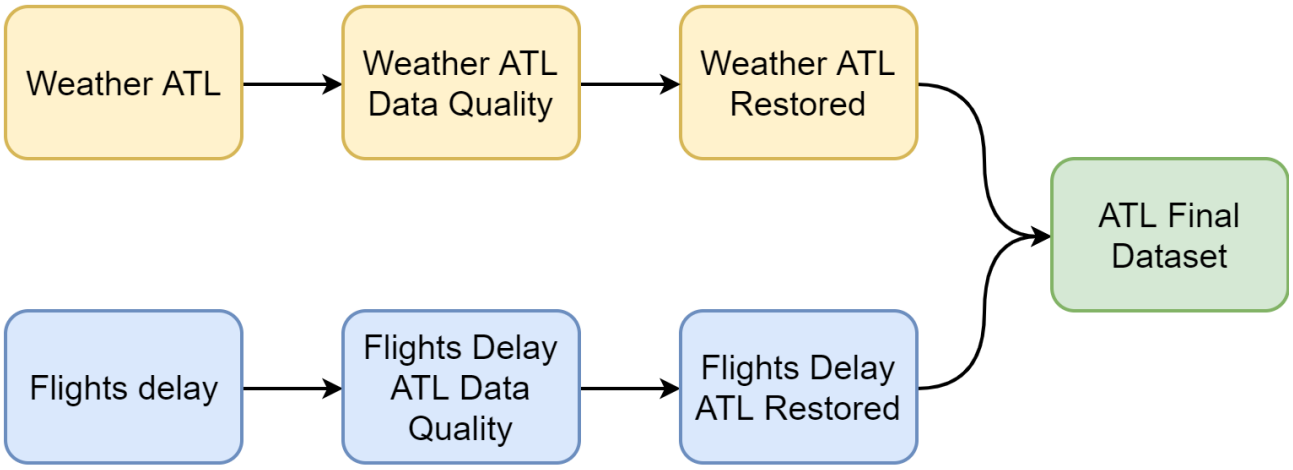
Nella fase successiva avverrà l'unione dei due dataset e la descrizione del dataset finale utilizzato per la risoluzione del problema di machine learning.

4 DATASET FINALE

4.1 Integrazione dei dataset

Ottenuti i dataset definitivi in ognuna delle due sezioni separate (le condizioni metereologiche e i voli aerei) si hanno ora tutti gli strumenti per costruire un dataset che contiene tutte e sole le istanze rilevanti, descritte con gli attributi selezionati, allo scopo di essere utilizzato per la risoluzione del problema di machine learning riguardo la classificazione dei voli aerei.

La creazione di questo nuovo e finale dataset *ATL_Final_Dataset* avviene attraverso il merge dei due dataset *Flights_delay_ATL_Restored* e *Weather_ATL_Restored* attraverso le colonne *"Day"* e *"Month"* assenti inizialmente nel dataset delle condizioni climatiche e aggiunte ad hoc per la fusione dei due dataset.



Nella prossima sezione sarà descritto nel dettaglio il dataset risultante.

4.2 Descrizione *ATL_Final_Dataset*

Questo database contiene 751636 istanze. Ciascuna istanza è relativa ad un volo aereo rappresentato con 15 attributi, 3 dei quali utili a rappresentare le condizioni metereologiche del giorno in cui è programmata la partenza.

Tutte e 15 derivano da uno degli atri due dataset, per questo non saranno descritti i dettagli tecnici che rimangono pressoché invariati per ogni attributo e possono essere consultati nella sezione 2.2. Sarà ora riportata solo una breve descrizione per ogni attributo in modo da avere chiari i dati che saranno poi usati nelle fasi successive.

Attributo	Descrizione
Month	Mese in considerazione.
Day	Giorno in considerazione.
Day of Week	Giorno della settimana preso in considerazione. 1 indica “lunedì”, 2 “martedì” ecc.
Airline	IATA Code della compagnia aerea.
Origin_Airport	IATA Code dell’aeroporto di partenza.
Destination_Airport	IATA Code dell’aeroporto di arrivo.
Scheduled_Departure	Orario di partenza pianificato.
Scheduled_Time	Tempo teorico trascorso dalla chiusura del gate dell’aeroporto di partenza fino all’arrivo al gate dell’aeroporto di destinazione.
Distance	Distanza in miglia tra l’aeroporto di partenza e quello di destinazione.
Scheduled_Arrival	Orario di arrivo pianificato.
Arrival_Time	Orario effettivo di arrivo.
Arrival_Delay	Minuti di ritardo o anticipo all’arrivo. È calcolato come: arrival_time - scheduled_arrival
MaximumTemp	Temperatura massima rilevata (° F).
MinimumTemp	Temperatura minima rilevata, (° F).
Precipitation	Precipitazioni medie, in inches all’ora.

4.3 Verifica della qualità dei dati

Creato il nuovo dataset si è immediatamente controllata la qualità dei dati, per accertarsi che la fase di merge sia stata effettuata correttamente.

Per fare ciò abbiamo nuovamente considerato le dimensioni utilizzate per la verifica della qualità dei due dataset iniziali. Nella tabella sottostante saranno riportati dimensioni, metriche e risultati applicati ad: *ATL_Final_Dataset*.

Dimensione	Tipo	Metrica	Risultati
Completezza	Completezza della tabella	Numero di valori non nulli / Numero totale di valori	11274540 / 11274540 = 1.000
Accuratezza	Accuratezza sintattica	Numero di valori corretti / Numero totale di valori	751636 / 751636 = 1.000
Consistenza	Integrità referenziale	Numero di valori consistenti / Numero totale di valori	751636 / 751636 = 1.000
	Consistenza dei dati	Numero di valori consistenti / Numero totale di valori	751636 / 751636 = 1.000
Unicità	Unicità delle tuple	Numero di tuple duplicate	751636 - 751636 = 0 tuple duplicate

Come possiamo vedere dai risultati risolvere i problemi sui singoli dataset prima dell'unione dei due si è rivelata una buona strategia e il dataset finale è corretto in tutte le dimensioni qualitative considerate senza dover eseguire alcuna elaborazione.

È possibile visionare il codice R per il controllo nel file "Data_Technology\Code\3_Data_Quality\ Data_Quality_Final_Dataset.R"

5 SCELTA DEL MODELLO

5.1 Modello selezionato

Durante la fase di analisi dei dataset e l'esplorazione di questi ultimi si è concluso che il modello migliore per la descrizione dei dati a disposizione fosse il modello relazionale. Questo perché i dati iniziali erano strutturati in modo tale che ogni oggetto del mondo reale fosse rappresentato da una riga con i diversi attributi utili a identificarlo, una chiave primaria univoca per quell'oggetto e una chiave esterna. Questa chiave esterna permette il collegamento tra le varie tabelle, divise per classi di oggetti. In particolare era presente una tabella per i voli, una per le compagnie aeree, una per gli aeroporti e l'ultima relativa alle condizioni meteo. Ogni tabella era rappresentata da un diverso file .csv.
Per questi motivi un modello relazionale è sembrato quello che potesse garantire una rappresentazione consistente (inserendo vincoli di integrità referenziale e sui dati) e logica (senza uno stravolgimento non giustificato dei dati iniziali) delle informazioni anche per il database finale. Una struttura simile è facilmente implementabile grazie ad un RDBMS. Essendo i dati in forma tabellare, però, possono anche essere trattati con un qualsiasi linguaggio di programmazione in grado di manipolare tabelle. Proprio per questo motivo, tutta la parte di manipolazione dei dati è stata implementata attraverso l'utilizzo di R.

5.2 Scelta di un possibile modello futuro

In una prima fase iniziale si è comunque considerata l'ipotesi di utilizzare un modello non relazione NoSQL e, in particolare, di modificare i dataset costruendo una nuova struttura document-based. La principale motivazione è che, innanzitutto è sembrato logicamente corretto raggruppare tutte le informazioni relative ad un oggetto del mondo reale che vogliamo rappresentare (un volo aereo) in un unico file e che, questo tipo di struttura dei dati fornisce strumenti simili a quelli che si possono usare con un modello relazionale.
In particolare, le tecniche di embedding e referencing possono aiutare a ricostruire le relazioni tra i vari oggetti evitando ripetizioni eccessive nei dati, mantendo chiara la struttura. Per fare questo, però, sarebbe stato necessario recuperare tutte le informazioni relative ad un volo aereo dalle tabelle (attraverso query SQL) e raggrupparle, strutturate, in un file JSON.

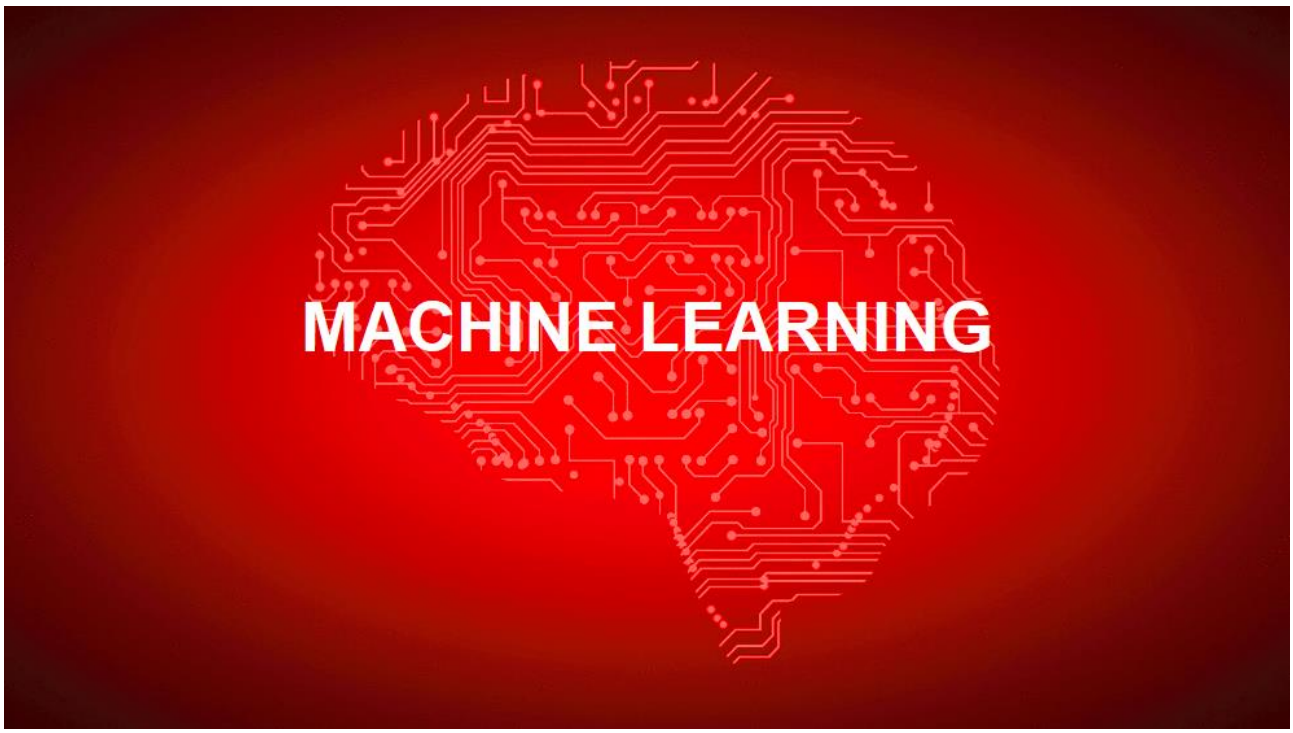
Se, ad esempio, si volesse trasformare un'istanza del dataset relazionale *Flights_ATL*:

Flights_ATL (*"AIRLINE" , "ORIGIN_AIRPORT" , "DESTINATION_AIRPORT , …*)

in un file JSON, i dati recuperati con una query e trasformati in un file JSON di questo tipo:

{*"AIRLINE":"NIK" , "ORIGIN_AIRPORT": "ATL" , "DESTINATION_AIRPORT": "TMP" , …*}

Questo processo, però, è stato reputato lento e complesso rispetto ai guadagni effettivi che avrebbe potuto portare al lavoro, considerando che non ci sarebbe dovuta essere una gestione futura dei dati già in nostro possesso, né di nuovi dati futuri.
Per la gestione di dati di questo tipo si sarebbe potuto utilizzare un DBMS non relazionale document based come MongoDB.


MACHINE LEARNING

6 SCELTA DEI DATI RILEVANTI

Prima di procedere con il lavoro vero e proprio per quanto riguarda la parte di machine learning, si è affrontata un'ulteriore fase di selezione dei dati rilevanti prendendo in considerazione il dataset generato.

Questa fase è stata necessaria, nonostante il database finale fosse decisamente più piccolo rispetto a quello iniziale, per problemi tecnici dovuti alla potenza di calcolo. *Flights_ATL* conteneva più di 7 milioni di voli, che sono poi stati ridotti a circa 750.000 nel dataset finale descritto nella fase "Final_Dataset_Description". Nonostante questo primo filtraggio delle istanze, si è deciso di ridurre ancora il numero di voli considerati riducendo il problema di machine learning non più alla determinazione del ritardo dei voli da e per l'aeroporto Hartsfield–Jackson Atlanta International Airport, ma di limitare i valori ad una sola tratta che, ovviamente, comprendeva anche questo aeroporto.

A questo punto sono state selezionate da *ATL_Final_Dataset* le tratte che contavano più istanze e per ogni sottoinsieme di valori è stato effettuato un piccolo studio col fine di capire quale dataset presentasse una maggiore correlazione tra gli attributi.

Per fare ciò è stata calcolata la correlazione tra le colonne a valori numerici dei vari dataset e l'attributo *"ARRIVAL_DELAY"* in modo da stimare a priori, innanzitutto quali colonne sarebbero state più utili nella predizione e, in secondo luogo, quale tratta aveva una maggior corrispondenza tra le colonne e il valore da classificare.

Queste correlazioni sono riportate, per ogni tratta considerata, nelle tabelle sottostanti.

	ATL - MCO						
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	1.000.000.000	0.805259452	0.11822792	0.157439043	0.003719629	0.003445125	-0.002831835
SCHEDULED_ARRIVAL	0.805259452	1.000.000.000	0.23971845	0.129300149	-0.006048975	-0.007136119	-0.002128444
SCHEDULED_TIME	0.118227925	0.239718451	100.000.000	-0.124153605	0.043152021	0.035895674	-0.015884996
ARRIVAL_DELAY	0.157439043	0.129300149	-0.12415360	1.000.000.000	0.006333677	0.065782110	0.158197251
MAXTEMP	0.003719629	-0.006048975	0.04315202	0.006333677	1.000.000.000	0.915046571	-0.023970313
MINTEMP	0.003445125	-0.007136119	0.03589567	0.065782110	0.915046571	1.000.000.000	0.112443119
PRECIPITATION	-0.002831835	-0.002128444	-0.01588500	0.158197251	-0.023970313	0.112443119	1.000.000.000

	ATL - LGA						
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	1.000.000.000	0.996981653	0.175844667	0.10461545	-0.003619272	-0.003419467	-0.007014545
SCHEDULED_ARRIVAL	0.996981653	1.000.000.000	0.212537338	0.10000852	-0.005503336	-0.004841698	-0.006635020
SCHEDULED_TIME	0.175844667	0.212537338	1.000.000.000	-0.07774165	-0.028421570	-0.023162915	0.005740007
ARRIVAL_DELAY	0.104615449	0.100008519	-0.077741651	100.000.000	-0.042153238	-0.023633749	0.104472376
MAXTEMP	-0.003619272	-0.005503336	-0.028421570	-0.04215324	1.000.000.000	0.919665258	-0.026661182
MINTEMP	-0.003419467	-0.004841698	-0.023162915	-0.02363375	0.919665258	1.000.000.000	0.105719167
PRECIPITATION	-0.007014545	-0.006635020	0.005740007	0.10447238	-0.026661182	0.105719167	1.000.000.000

ATL - FLL							
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	1.000.000.000	0.765120590	0.12464231	0.15770898	0.002286905	0.003260186	-0.001667385
SCHEDULED_ARRIVAL	0.765120590	1.000.000.000	0.24350277	0.12646016	-0.031995648	-0.031132094	-0.003042346
SCHEDULED_TIME	0.124642308	0.243502767	100.000.000	-0.10341754	-0.025112742	-0.046949435	-0.025025409
ARRIVAL_DELAY	0.157708981	0.126460162	-0.10341754	100.000.000	-0.014566089	0.039892703	0.153002414
MAXTEMP	0.002286905	-0.031995648	-0.02511274	-0.01456609	1.000.000.000	0.916146599	-0.020720398
MINTEMP	0.003260186	-0.031132094	-0.04694943	0.03989270	0.916146599	1.000.000.000	0.114213843
PRECIPITATION	-0.001667385	-0.003042346	-0.02502541	0.15300241	-0.020720398	0.114213843	1.000.000.000

ATL - TPA							
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	10.000.000.000	0.785273732	0.034439609	0.16361372	0.027477109	0.027027794	-0.0005467464
SCHEDULED_ARRIVAL	0.7852737323	1.000.000.000	0.162905791	0.14293646	-0.003093633	-0.002325459	0.0013853632
SCHEDULED_TIME	0.0344396094	0.162905791	1.000.000.000	-0.11098885	-0.008862007	-0.019197726	-0.0203886819
ARRIVAL_DELAY	0.1636137171	0.142936460	-0.110988845	100.000.000	0.011465463	0.067327200	0.1694035843
MAXTEMP	0.0274771085	-0.003093633	-0.008862007	0.01146546	1.000.000.000	0.913887961	-0.0237964963
MINTEMP	0.0270277943	-0.002325459	-0.019197726	0.06732720	0.913887961	1.000.000.000	0.1128834168
PRECIPITATION	-0.0005467464	0.001385363	-0.020388682	0.16940358	-0.023796496	0.112883417	10.000.000.000

ATL - DFW							
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	1.000.000.000	0.938918162	0.112471823	0.098484497	0.02572783	0.028505609	0.001178289
SCHEDULED_ARRIVAL	0.938918162	1.000.000.000	-0.015808994	0.101405889	0.02746647	0.028484177	-0.004975269
SCHEDULED_TIME	0.112471823	-0.015808994	1.000.000.000	-0.070027934	0.02168976	0.020836509	0.001400887
ARRIVAL_DELAY	0.098484497	0.101405889	-0.070027934	1.000.000.000	-0.02821309	0.005753398	0.171065415
MAXTEMP	0.025727834	0.027466474	0.021689760	-0.028213089	100.000.000	0.916496713	-0.021155726
MINTEMP	0.028505609	0.028484177	0.020836509	0.005753398	0.91649671	1.000.000.000	0.110717212
PRECIPITATION	0.001178289	-0.004975269	0.001400887	0.171065415	-0.02115573	0.110717212	1.000.000.000

ATL - PHL							
	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	SCHEDULED_TIME	ARRIVAL_DELAY	MAXTEMP	MINTEMP	PRECIPITATION
SCHEDULED_DEPARTURE	1.000.000.000	0.583766365	0.06604416	0.10509200	0.006638944	0.006043442	-0.002069349
SCHEDULED_ARRIVAL	0.583766365	1.000.000.000	0.18682803	0.07943884	-0.064254124	-0.069562941	-0.003665323
SCHEDULED_TIME	0.066044162	0.186828033	100.000.000	-0.04599004	-0.038409066	-0.041542119	-0.011265243
ARRIVAL_DELAY	0.105092000	0.079438844	-0.04599004	100.000.000	0.017648473	0.073954682	0.164206363
MAXTEMP	0.006638944	-0.064254124	-0.03840907	0.01764847	1.000.000.000	0.916588580	-0.021719917
MINTEMP	0.006043442	-0.069562941	-0.04154212	0.07395468	0.916588580	1.000.000.000	0.110248731
PRECIPITATION	-0.002069349	-0.003665323	-0.01126524	0.16420636	-0.021719917	0.110248731	1.000.000.000

Come possiamo notare, calcolando la correlazione con il coefficiente di Pearson, non si ottengono valori superiori a 0,171. Questo è dovuto al fatto che, essendo un problema reale non è così ovvio che ci sia una grande correlazione tra il valore degli attributi selezionati e l'effettivo ritardo. Infatti, scegliendo di considerare solo le informazioni che si conoscono a priori, sono stati scartati molti dati che probabilmente avrebbero reso la classificazione molto più semplice e precisa ma che, al contempo, avrebbero reso il problema meno realistico e più banale.

Per la scelta definitiva della tratta da utilizzare è stato assegnato un valore da 0 a 2 (con mezzi punti) alle correlazioni riscontrate nelle varie tabelle, con i risultati riportati di seguito.

		AIRPORT IATA_CODE					
		MCO	LGA	FLL	TPA	DFW	PHL
ATTRIBUTE	DAY	0	1	0	1*	0	1*
	DAY_OF_WEEK	0	1*	0	0	0	0
	MONTH	0	1*	0	1	0	0
	DEST_AIRPORT	0	0	0	0	0	0
	AIRLINE	1	0	1*	2	0	1*

0	Bad
1	Medium
2	Good
*	Half value

AIRPORT SELECTED	MCO	LGA	FLL	TPA	DFW	PHL
	X	X	X	V	X	X

La tratta selezionata è quindi risultata essere “ATL” – “TMP”. Il dataset finale per il machine learning è quindi TPA_Dataset, ovvero un sottoinsieme di ATL_Final_Dataset in che mantiene invariati gli attributi ma, considera solamente le 15350 tuple che riportano “ATL” o “TMP” negli attributi che riguardano l’aeroporto di arrivo e di destinazione.

7 ANALISI ESPORATIVA TRAINING SET

7.1 Scelta training set

Una volta costruito il dataset da utilizzare, si è passati alla creazione del training set e del test set. Si è deciso di dividere *TPA_Dataset* in due parti: il 70% come training set e il 30% come test set. Nel fare ciò, però, si è prestata molta attenzione alla distribuzione dei ritardi e, infatti, si è cercato di fare in modo che *TPA_Dataset, TPA_Trainingset* e *TPA_Testset* avessero le stesse percentuali di voli in ritardo. Lo scopo di questa operazione è quello di ottenere training set che fosse il più possibile rappresentativo dell’intero database.

7.2 Esporazione dei dati

Ottenuto il training set è stata condotta un’analisi esplorativa dei dati in esso contenuti, effettuando alcuni test per comprendere meglio eventuali correlazioni tra dati e poter effettuare con maggior precisione ed efficacia le successive fasi.

Tutte le statistiche sono state effettuate utilizzando R ed è possibile visionare il file in “Machine_Learning\Code\2_Data_Exploration \Data_Exploration.R”

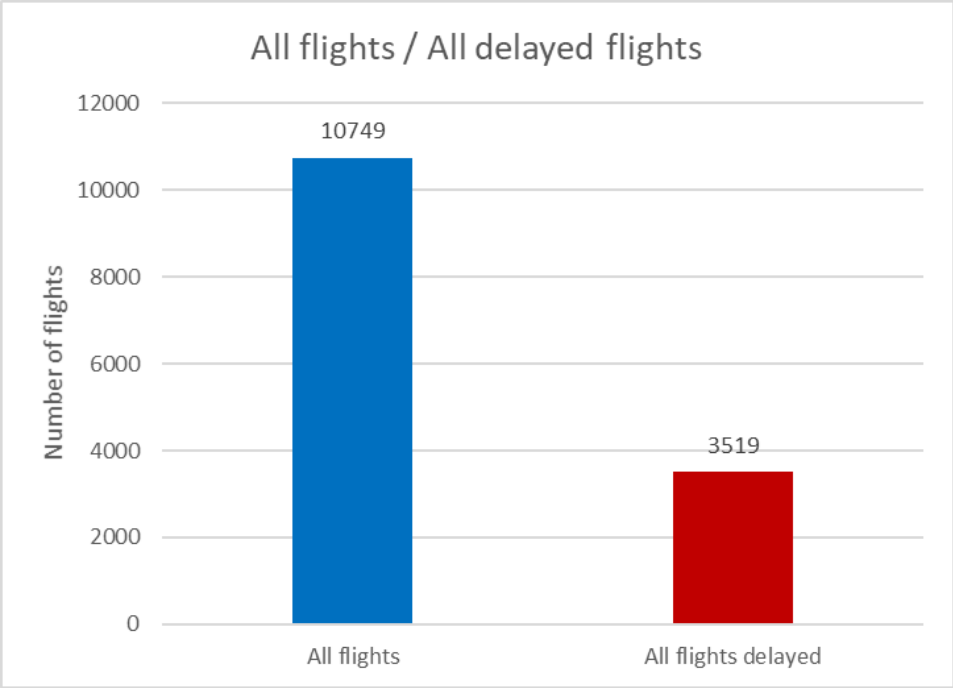
7.2.1 Dataset *TPA_Trainingset*

Il training set contiene il 70% delle 15350 istanze del dataset *TPA_Dataset*, ovvero 10749 righe.

All flights / All delayed flights

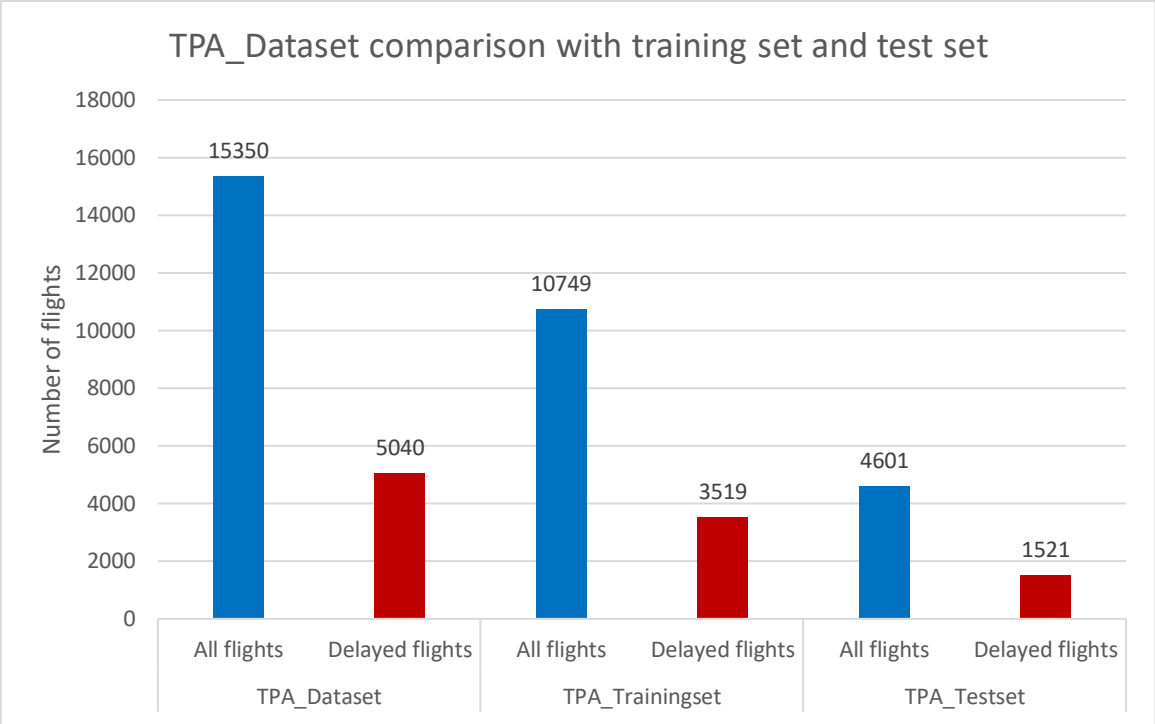
Analizzando la distribuzione dei voli nel dataset è stato visualizzato il numero totale dei voli e calcolato quello totale dei voli in ritardo in modo da ottenere una proporzione tra questi due valori.
I risultati sono mostrati nella tabella e nel grafico qui riportati.

All flights	All flights delayed
10749	3519



Percentage of deleyed flights
33%

Come si può notare il rapporto tra i voli totali e i voli in ritardo è del 33%. Il dataset *ATL_Final_Dataset* del 31%, perciò essi hanno una distribuzione simile che ci permette di lavorare con dati proporzionati. Inoltre, come accennato nell'introduzione a questa parte, si è cercato di mantenere costante la distribuzione dei ritardi tra i vari dataset riguardanti la tratta "*ATL*" – "*TMP*". I dati sono riportati nei grafici sottostanti e, come si può notare, tutti i dataset condividono il 33% dei ritardi.



TPA_Dataset	
All flights	Delayed flights
15350	5040
Percentage of delayed flights	
33%	

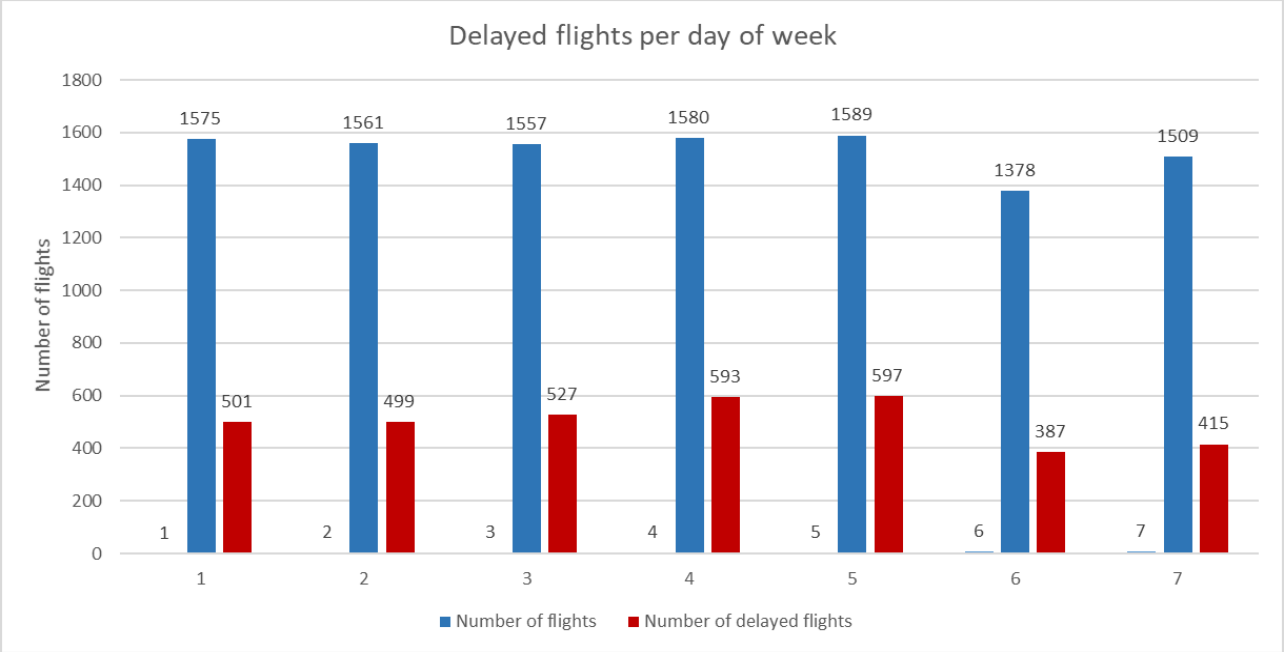
TPA_Trainingset	
All flights	Delayed flights
10749	3519
Percentage of delayed flights	
33%	

TPA_Testset	
All flights	Delayed flights
4601	1521
Percentage of delayed flights	
33%	

Delayed flights per day of week

A questo punto si è cercata una correlazione tra i ritardi e il giorno della settimana in cui è partito il volo, ottenendo i dati qui riportati.

Days of week	Number of flights	Number of delayed flights
1	1575	501
2	1561	499
3	1557	527
4	1580	593
5	1589	597
6	1378	387
7	1509	415



Percentage of delayed flights per day						
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
32%	32%	34%	38%	38%	28%	28%

Come si può notare dai dati ottenuti i voli maggiormente in ritardo partono giovedì e venerdì, mentre quelli meno in ritardo sabato e domenica.

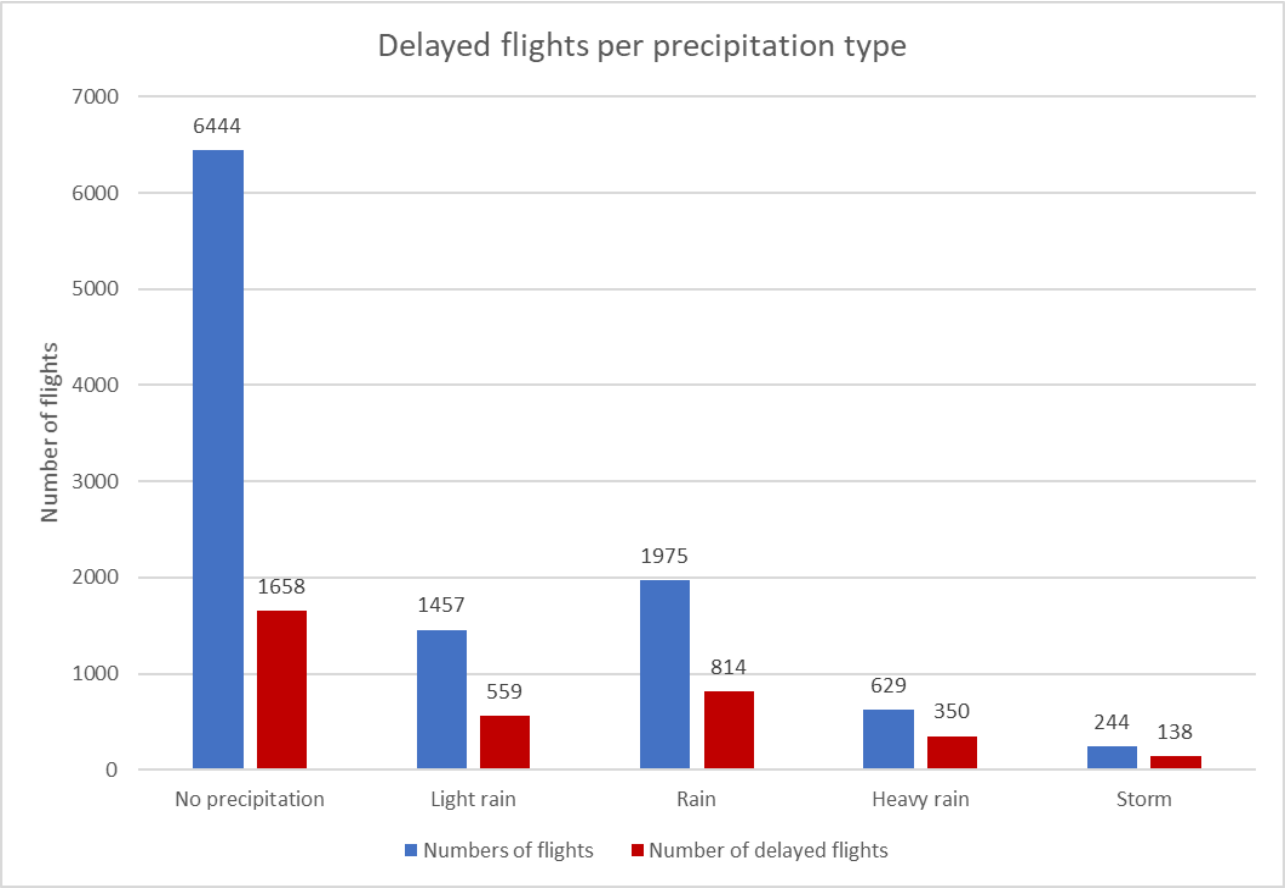
Delayed flights per precipitation

In questa fase abbiamo deciso di dividere i valori relativi alle precipitazioni nei 5 gruppi precedentemente utilizzati per l'analisi del dataset *Weather_ATL* nella sezione 2.4.2, calcolando il numero totale dei voli e i voli in ritardo a seconda delle precipitazioni.

Inches per hour				
No precipitation	Light rain	Rain	Heavy Rain	Storm
0	$0 < x \leq 0{,}1$	$0{,}1 < x \leq 0{,}79$	$0{,}79 < x \leq 1{,}5$	$x > 1{,}5$

I risultati ottenuti sono riportati qui di seguito.

Precipitation type	Numbers of flights	Number of delayed flights
No precipitation	6444	1658
Light rain	1457	559
Rain	1975	814
Heavy rain	629	350
Storm	244	138



Come si evince dalla tabella precedente, c'è una correlazione rilevante tra le precipitazioni e i ritardi dei voli: man mano che le condizioni metereologiche relative alle precipitazioni peggiorano (aumenta l'intensità della pioggia), maggiori saranno i ritardi. Infatti, in caso di tempesta o forti piogge i voli ritarderanno rispettivamente del 57% e 56%

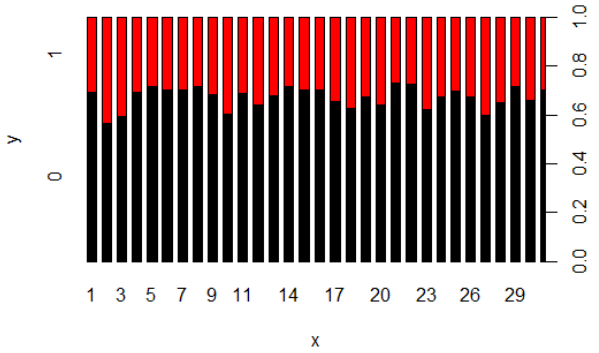
Mentre, ad esempio, senza precipitazioni il ritardo medio è solamente del 26%.
Quindi con condizioni di precipitazioni avverse il numero dei voli in ritardo sarà più del doppio rispetto a quanto non sono previste piogge.

Percentage delayed flights per precipitation				
No precipitation	Light rain	Rain	Heavy rain	Storm
26%	38%	41%	56%	57%

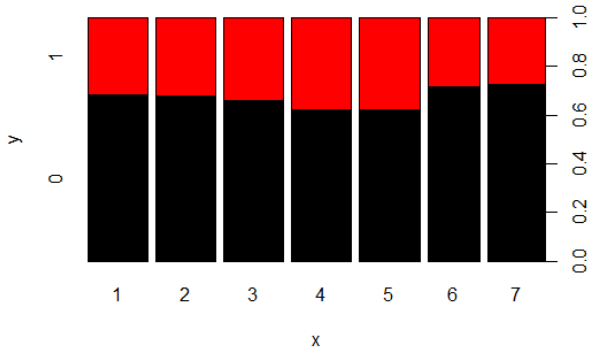
7.3 Grafici delle correlazioni Attributo - Ritardo

In seguito sono elencati i grafici, mettendo in relazione alcuni attributi rilevanti con il ritardo del volo (“*ARRIVAL_DELAY”*).
In ognuno dei grafici i voli in ritardo sono rappresentati seguendo la seguente legenda.

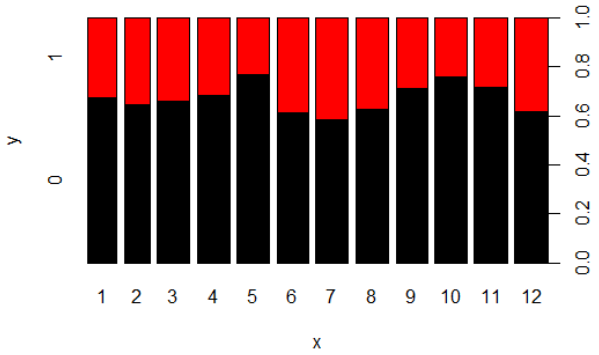
Day – Arrival_Delay



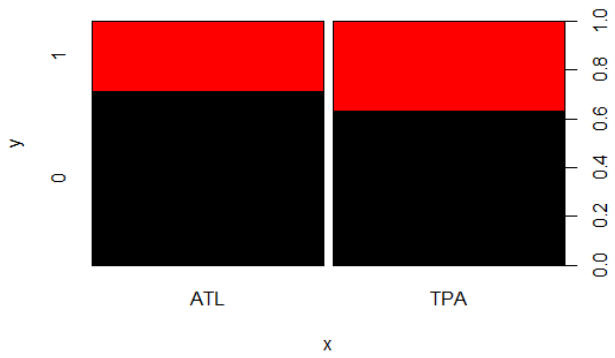
Day_of_Week – Arrival_Delay



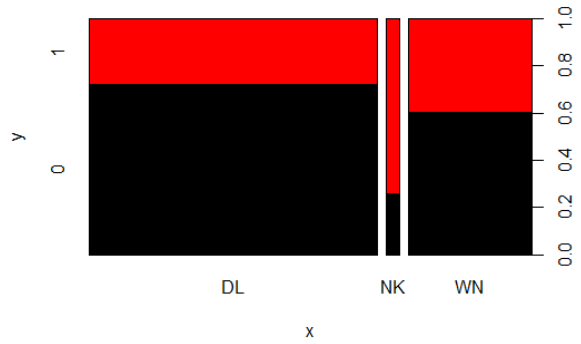
Month – Arrival_Delay



Destination_Airport – Arrival_Delay



Airline – Arrival_delay



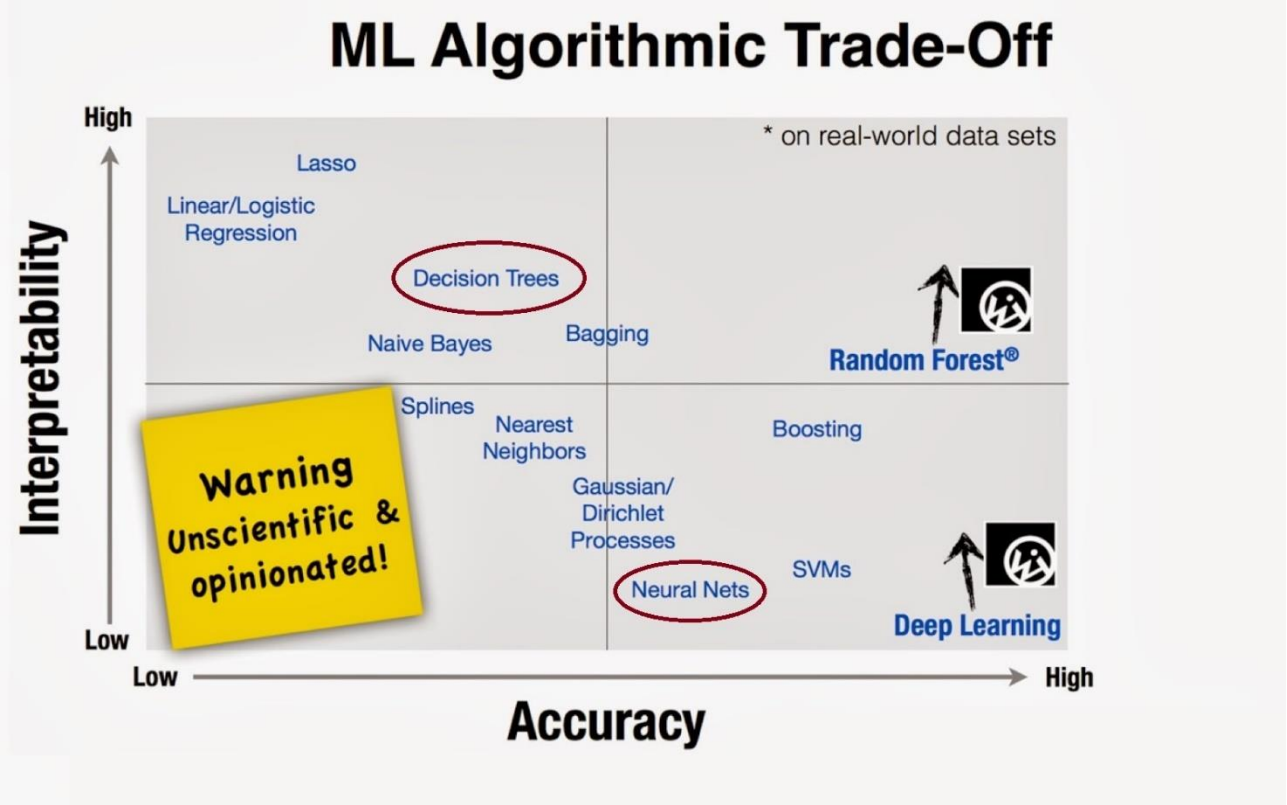
Come si può notare in tutti i grafici precedentemente mostrati la correlazione con gli attributi analizzati e il ritardo è poco significativa, fatta eccezione per l’attributo *“AIRLINE”*. Infatti in questo caso è evidenziato un netto ritardo della compagnia aerea NK, rispetto alle altre due per compagnie che abitualmente viaggiano in questa tratta (DL e WN).

8 ANALISI DEI MODELLI DI MACHINE LEARNING UTILIZZATI

8.1 Scelta dei modelli

Dopo un'analisi esplorativa del dataset *TPA_Trainingset*, si è cercato di capire con quale tecnica di machine learning si sarebbero ottenuti i risultati migliori.

La prima riflessione riguarda il significato di "migliori": le due dimensioni che maggiormente possono modificare il significato di questo termine sono la precisione della classificazione e l'interpretabilità dei dati. Come presentato nell'immagine sottostante queste due dimensioni sono complementari: all'aumentare dell'interpretabilità, diminuiscono le performance e viceversa.



Inizialmente, un modello che avesse un'alta interpretabilità dei dati era necessario per comprendere quali attributi fossero più significativi per la classificazione. Proprio per questo motivo si è deciso di utilizzare un albero di decisione.

I risultati dell'albero di decisione, sono poi stati utilizzati per decidere i parametri in input di un modello più complesso e performante: una rete neurale.

Saranno descritti nelle prossime sezioni i modelli utilizzati e i ragionamenti che hanno portato al passaggio da un modello ad all'altro.

8.2 Decision Tree

8.2.1 Descrizione del modello

Dopo una prima fase di studio teorico riguardo gli alberi di decisione si è deciso di iniziare a testare un albero utilizzando tutti gli attributi presenti in *TPA_Trainingtest*.

Inizialmente il problema si è rivelato quello della scelta dei valori dei parametri per ottimizzare il modello secondo gli scopi del lavoro. In particolare i parametri considerati sono descritti di seguito.

8.2.2 Soglia di assegnazione

L'albero di decisione restituisce un valore numerico per ogni istanza data in input alla fase di testing. L'output di default è un valore compreso tra 0 e 1:

- Ai valori compresi tra 0 e 0.5 è associata la label 0 (in orario)
- AI valori compresi tra 0.5 e 1 è associata la label 1 (in ritardo)

La modifica di questa soglia permette di gestire la categorizzazione delle istanze in input. La motivazione che giustifica questo cambio di soglia è che raramente i valori restituiti dall'albero in fase di testing delle istanze erano compresi tra 0 e 1: molto spesso il valore massimo era più basso e, per questo motivo, si è deciso di cercare un corretto valore per questa soglia.

8.2.3 CP

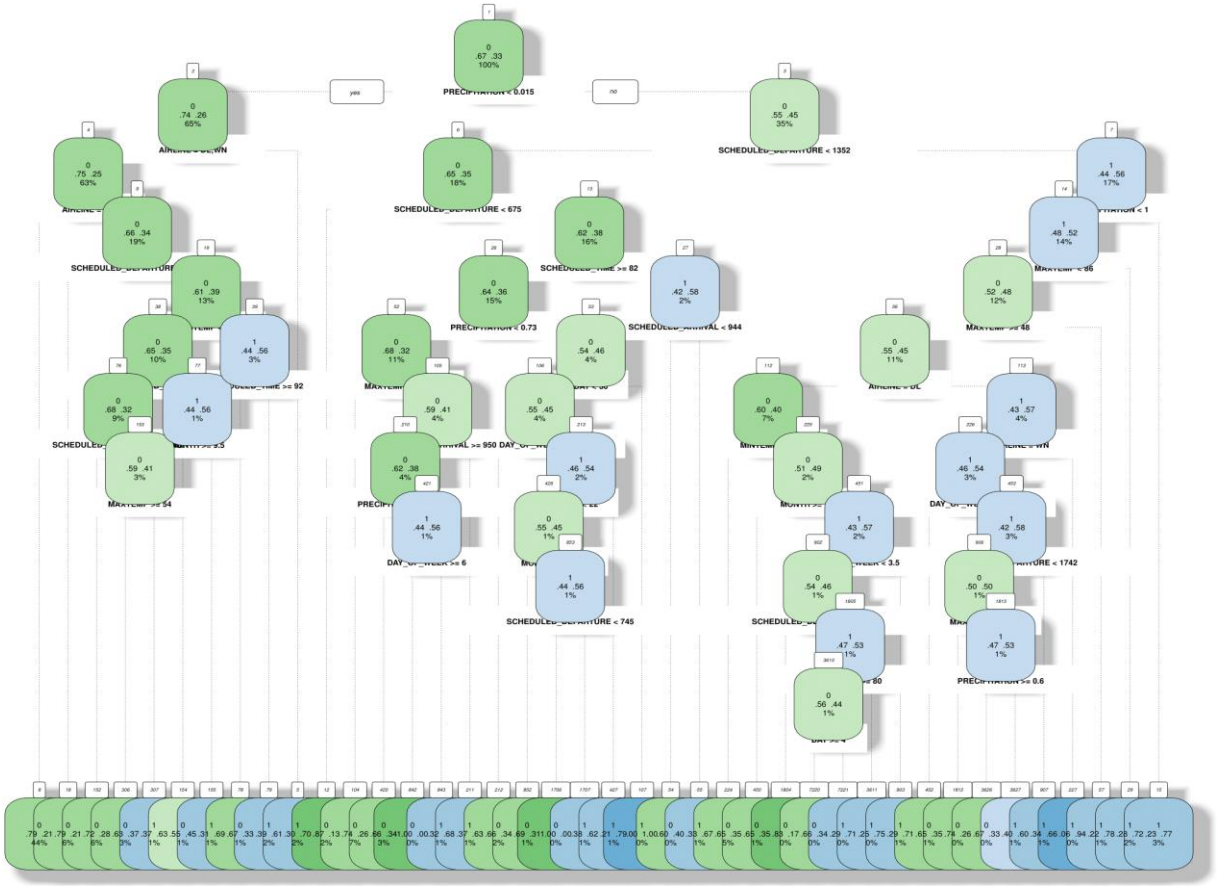
Il parametro di complessità dell'albero ne gestisce il livello di potatura e permette di avere un albero più o meno profondo.

Capire dove potare l'albero non solo permette di migliorare le performance, ma garantisce anche una maggior leggibilità e interpretabilità della struttura del classificatore.

Lo scopo del primo test era quello di capire quali attributi fossero più importanti ai fini della classificazione e quale fosse una soglia corretta per classificare al meglio le istanze più rilevanti ai fini del lavoro, ovvero i voli in ritardo (True_NEG).

8.2.4 Scelta del modello e performance evaluations

Il primo albero è stato testato con il valore standard di soglia e un livello molto basso di potatura, sulla totalità degli attributi di *TPA_Trainingset*. Il risultato è quello riportato nell'immagine sottostante.



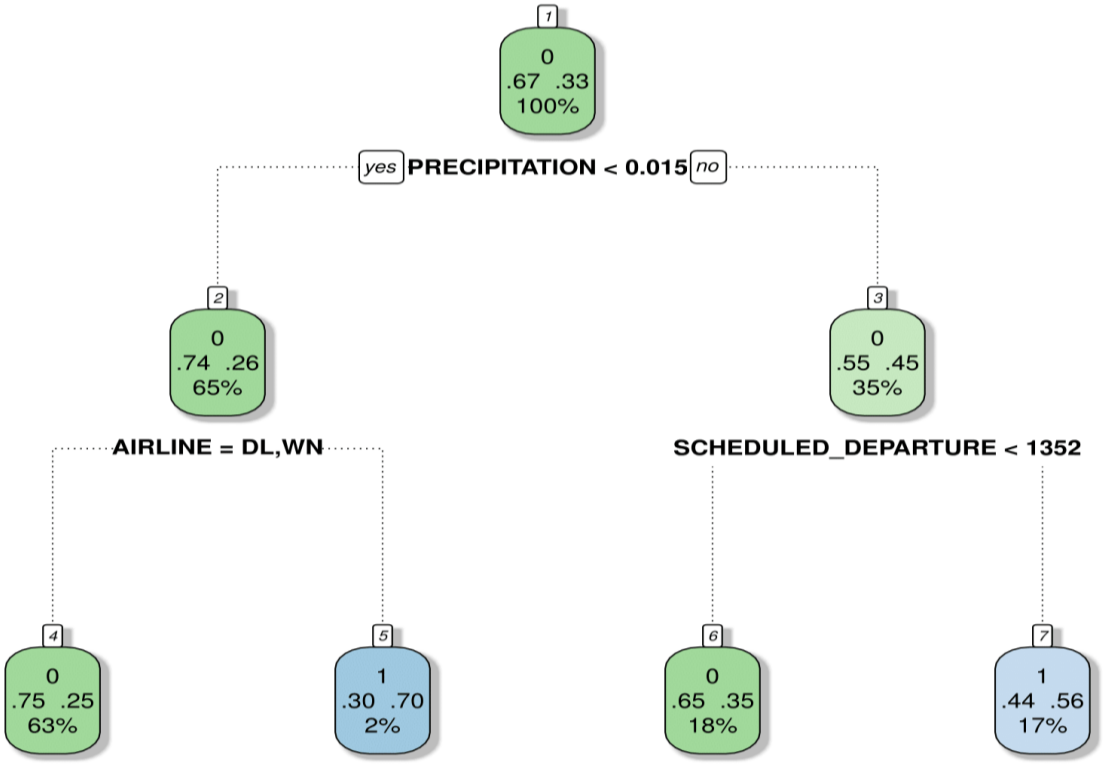
L'albero risultava troppo grande per essere compreso e, per questo, è stato nuovamente testato aumentando il fattore di taglio (CP), per migliorarne la leggibilità, e una soglia di assegnazione più bassa. Quest'ultimo passaggio ha come scopo la massimizzazione dei True_NEG minimizzando la perdita di True_POS. Questo è necessario al fine di ottenere un modello che si adatti meglio al nostro problema perché, non sempre, la corretta classificazione di tutte le label ha la stessa importanza. Nel nostro caso, identificare i voli in orario (True_POS) in un dataset in cui la percentuale di questi ultimi è doppia rispetto a quella di voli in ritardo, ha una minor importanza rispetto alla classificazione corretta dei voli in ritardo (True_NEG).

Questa operazione è stata iterata due volte con i dati riportati nella tabella sottostante.

	Config #1	Config #2	Config #3
CP	0.0016	0.0066	0.016
Threshold	0.5	0.4	0.3
Variables	13	13	13
Mean_Acc	0.681	0.686	0.675
True_POS	2798	2898	2117
False_NEG	282	182	963
False_POS	952	1139	702
True _NEG	569	382	819
F_Measure_0	0.819	0.814	0.718
F_Measure_1	0.48	0.366	0.496
AUC	0.6413	0.5960	0.6129

Il grafico dell'albero della configurazione #3, che possiamo vedere nell'immagine sottostante, risulta adeguatamente potato e con una soglia adeguata al tipo di risultati che si possono richiedere ad un classificatore di questo tipo. Come possiamo notare, però, il nuovo albero non utilizza tutti e 13 gli attributi su cui è stato per la classificazione ma ne utilizza solamente 3.

A questo punto è sembrato naturale cercare di ottenere gli stessi risultati trainando un modello che, a priori, utilizzi i soli attributi *"Precipitation"*, *"Scheduled_Departure" e "Airline".* I risultati ottenuti anche in questo nuovo training sono effettivamente gli stessi.



Questo si può notare guardando i valori della matrice di confusione: nelle configurazioni #3, #5 e #6, infatti, i valori sono gli stessi. La struttura dell'albero finale è quindi quella riportata sopra. La valutazione delle performance di quest'ultima configurazione è riportata nel paragrafo "Performance evaluations".

	Config #4	Config #5	Config #6
CP	0.0016	0.0066	0.016
Threshold	0.3	0.3	0.3
Variables	3	3	3
Mean_Acc	0.661	0.672	0.675
True_POS	1871	2117	2117
False_NEG	1209	963	963
False_POS	481	702	702
True _NEG	1040	819	819
F_Measure_0	0.689	0.718	0.718
F_Measure_1	0.552	0.496	0.496
AUC	0.6456	0.6129	0.6129

8.3 Neural Network

8.3.1 Descrizione del modello

Considerando i risultati ottenuti dal modello dell'albero di decisione, si è deciso di costruire un modello di rete neurale rimanendo coerenti con il numero ed il tipo di variabili utilizzate, ovvero:

- *"Scheduled_Departure"*
- "*Airlines*"
- "Precipitation"

Come prima cosa il modello è stato eseguito con i parametri di default, dopo di che ci si è posti il problema di scelta dei valori da associare a ciascuno di essi in modo da cercare di ottimizzare il modello per il problema di classificazione corrente. I parametri presi in considerazione sono stati:

8.3.2 Numero di neuroni

Ci sono infinite teorie che riportano possibili soluzioni al problema di decisione del numero di livelli e di neuroni per trainare una rete neurale; nonostante ciò è tutt'ora difficile riuscire a trovare una regola ben definita per un problema specifico.
Per questo motivo durante questa fase del progetto la scelta del numero di neuroni ha acquisito una discreta rilevanza, portando quindi ad una fase di testing e valutazione delle possibili configurazioni utilizzabili per il nostro caso specifico.

8.3.3 Soglia minima di arresto

Questo parametro permette di modificare la soglia di errore complessivo del modello; ciò avviene forzando la fase di training a continuare a cercare una soluzione migliore fino al raggiungimento della soglia data in input.
Durante questa fase, è stato ovviamente considerato il fatto che all'abbassarsi di questa soglia corrisponde un aumento esponenziale del tempo richiesto per il training del modello, ma non sempre questo implica anche un aumento significativo delle performance.

8.3.4 Soglia di assegnazione

L'output di una rete neurale corrisponde ad un valore numerico associato ad ogni singola istanza data in input. Il valore dell'output è di default un valore compreso tra 0 e 1:

- Ai valori compresi tra 0 e 0.5 è associata la label 0 (in orario)
- Ai valori compresi tra 0.5 e 1 è associata la label 1 (in ritardo)

Come già riportato per l'albero di decisione, la modifica di questa soglia è giustificata dalla non corretta distribuzione dei valori in output nel range di valori tra 0 e 1.
Come si può notare, le label rispetto all'albero di decisione sono invertite.

8.3.5 Scelta della configurazione

Per cercare di comprendere al meglio il funzionamento di una rete neurale, durante la fase di scelta della configurazione del modello ne sono state testate svariate, di cui, le più rilevanti, sono riportate nella tabella sottostante.

	Config #1	Config #2	Config #3	Config #4	Config #5	Config #6	Config #7
Neurons_number	c(10,2)	25	20	15	10	5	3
Threshold	0,35	0,35	0,35	0,35	0,35	0,35	0,35
NN_Min_Threshold	0,5	0,25	0,25	0,01	0,25	0,01	0,01

Dopo aver confrontato i risultati ottenuti dai test delle varie configurazioni, si è giunti alla conclusione che nessuna delle configurazioni riporta delle performance nettamente migliori delle altre, ma ciò ha permesso di riportare alcune osservazioni rilevanti, in particolare:

- Considerando un singolo livello, all'aumentare del numero di neuroni, si è notato un lieve miglioramento delle performance.
- All'aumentare del numero di livelli di neuroni si è notato che, nonostante la potenza computazionale necessaria al training del modello aumentava esponenzialmente, le performance si sono rivelate presso che le stesse.

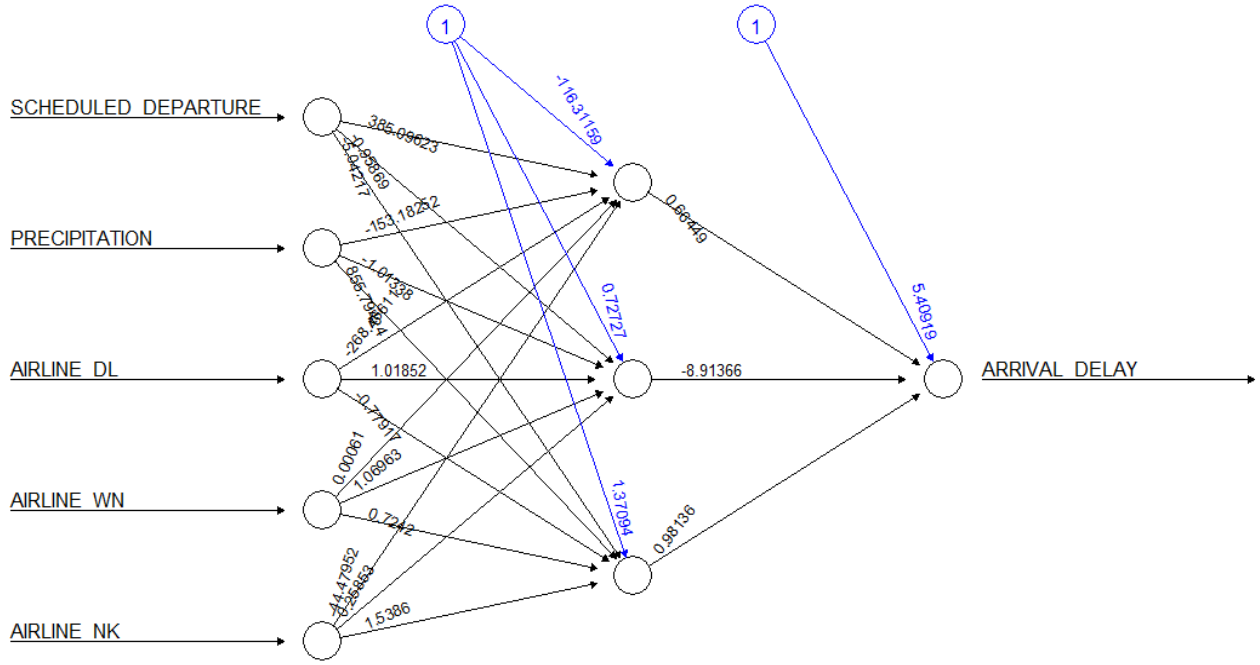
A questo punto si è presentato il problema di individuare la configurazione che permettesse di generare un modello in grado di classificare correttamente il maggior numero di voli in ritardo possibile (True_NEG), garantendo allo stesso tempo la minimizzazione sia della perdita di voli correttamente classificati in orario (True_POS) sia della mole computazionale richiesta per il training della rete neurale.

Nella tabella sottostante sono riportate le configurazioni testate che hanno riportato i migliori risultati numerici riguardanti le performance measures. In particolare, la configurazione #7 permette di ottenere un modello che più rispecchia le priorità sopra riportate, nonostante utilizzi solamente 3 neuroni.
E' possibile infatti notare come, nonostante il valore dell'AUC non sia il massimo riscontrato, tutti gli altri parametri di valutazione presentano valori in media migliori anche delle reti trainate con un maggior numero di neuroni. La scelta di bilanciamento tra voli correttamente identificati in ritardo e in orario è rimasta coerente con quella utilizzata nell'albero di decisione.

	Config #1	Config #2	Config #3	Config #4	Config #5	Config #6	Config #7
Neurons_number	c(10,2)	25	20	15	10	5	3
Threshold	0,35	0,35	0,35	0,35	0,35	0,35	0,35
NN_Min_Threshold	0,5	0,25	0,25	0,01	0,25	0,01	0,01
Mean_Acc	*	*	*		*		
True_POS	2180	2191	2207	2201	2168	2233	2137
False_POS	633	628	640	641	646	661	611
False_NEG	900	889	873	879	912	847	943
True_NEG	888	893	881	880	875	860	910
F_Measure_0	74,0%	74,3%	74,5%	74,3%	73,6%	74,8%	73,3%
F_Measure_1	53,7%	54,1%	53,8%	53,7%	52,9%	53,3%	53,9%
AUC	0,6458	0,6492	0,6479	0,6466	0,6396	0,6452	0,6461

* Il valore numerico di questo parametro non è stato calcolato a causa della mole computazionale richiesta

Di seguito è riportata la struttura del modello di rete neurale selezionato.



La valutazione del modello sarà riportata più nello specifico all'interno della sezione "Performance evaluations".

9 PERFORMANCE EVALUATIONS: NN & DT

Dopo aver scelto le configurazioni definitive per i due modelli, si è passati alla fase di valutazione delle performance di questi ultimi. Prima di procedere con l'analisi dei risultati ottenuti da questa fase, è indispensabile fare una premessa sulle metriche considerate per la valutazione di un modello.

Tutti i modelli sono stati valutati dando un diverso peso alla corretta identificazione delle due classi ("in orario" , "in ritardo"). Considerando che lo scopo principale del nostro problema di classificazione è quello di individuare correttamente il più alto numero di ritardi possibile, questa classe è decisamente più rilevante rispetto a quella a cui appartengono i voli in orario. Inoltre, le percentuali di istanze relative alle due classi sono molto diverse tra loro: il 67% dei voli all'interno del dataset risulta "in orario", mentre soltanto il 33% dei voli risulta "in ritardo". Proprio per questi motivi si è deciso di prioritizzare le misure di performance che riguardano la corretta identificazione delle istanze "in ritardo" (Precision_1, Recall_1, F_Measure_1).

Le tabelle di seguito permettono di confrontare i valori delle performance measures ottenute dai test eseguiti sui due modelli. Si ricorda che:

True_POS: voli realmente "in orario" correttamente classificati "in orario"
False_POS: voli "in ritardo" erroneamente classificati "in orario"
False_NEG: voli "in orario" erroneamente classificati "in ritardo"
True_NEG: voli realmente "in ritardo" correttamente classificati "in ritardo"

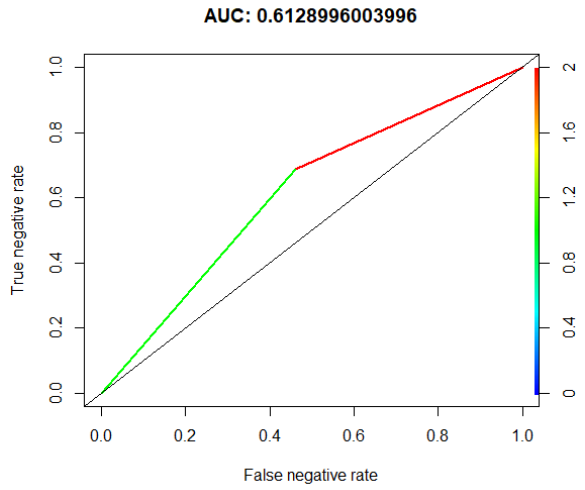
Decision Tree Model Configuration	
CP	0.016
Threshold	0.3
Variables	3
Testset: voli in orario	3080
Testset: voli in ritardo	1521
10-Fold cross validation	
Min_Acc	0.602
Mean_Acc	0.684
Max_Acc	0.707
Performance Evaluation	
Model_Acc	0.638
True_POS	2117
False_POS	702
False_NEG	963
True_NEG	819
Precision_0	0.751
Precision_1	0.460
Recall_0	0.687
Recall_1	0.538
F_Measure_0	0.718
F_Measure_1	0.496
AUC	0.6129

Neural Network Model Configuration	
Neurons.number	3
Threshold	0,35
NN_Min_Threshold	0,01
Testset: voli in orario	3080
Testset: voli in ritardo	1521
10-Fold cross validation	
Min_Acc	0.657
Mean_Acc	0.663
Max_Acc	0.667
Performance Evaluation	
Model_Acc	0.662
True_POS	2137
False_POS	611
False_NEG	943
True_NEG	910
Precision_0	0,778
Precision_1	0,491
Recall_0	0,694
Recall_1	0,598
F_Measure_0	0,733
F_Measure_1	0,539
AUC	0,6461

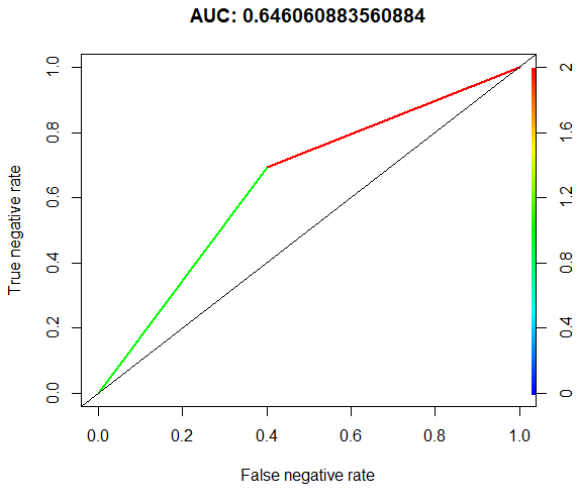
Come si può vedere, i valori assoluti riportati dagli AUC non sono ottimali. Questo, come tutti i risultati ottenuti durante la fase di valutazione delle performance, è dovuto alla bassa correlazione tra gli attributi del dataset e la relativa classe di appartenenza dell'istanza. Si è scoperto fin dalle prime fasi del progetto di questa bassa correlazione, ma, quando si trattano problemi reali, capita di rado di individuare una forte e chiara correlazione tra dati.

Sarebbe stato possibile ottenere valori di AUC maggiori prioritizzando la classificazione dei voli "in orario", ma un risultato del genere avrebbe avuto una minor rilevanza per il problema in questione, quindi si è deciso di trovare ed utilizzare delle configurazioni che permettessero di ottenere delle misure di performance che prioritizzassero la corretta classificazione delle istanze "in ritardo".
Infatti i grafici ROC che si possono visualizzare di seguito non riportano il ratio di true positives e false positives, bensì quello di true negatives e false negatives.

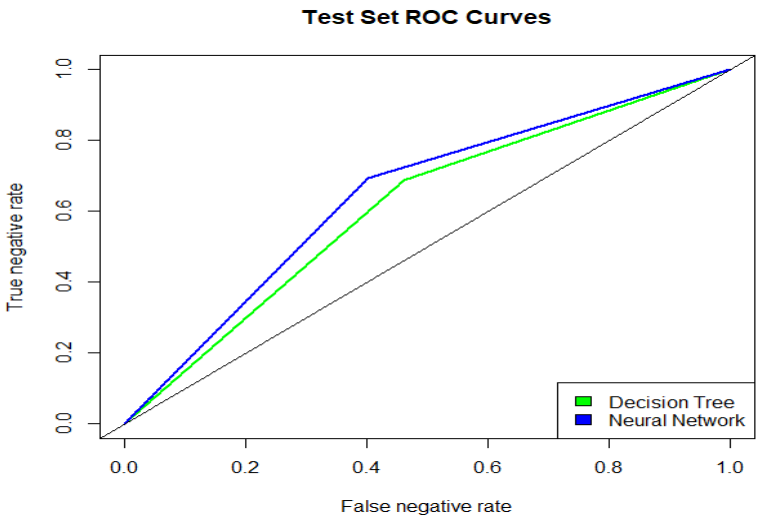
Decision Tree



Neural Network



Considerando le altre statistiche nelle tabelle si può notare che il modello di rete neurale implementato riporta risultati migliori riguardo a tutte le misure di performance considerate. Questo è dovuto dal fatto che, comparato con il modello di albero di decisione, il modello di rete neurale è in grado di classificare correttamente un maggior numero sia di voli "in orario" sia di voli "in ritardo". Inoltre, il grafico ROC riportato di seguito provvede un'ulteriore conferma riguardante sia il tasso di true negative sia il tasso di false negative.



Un ulteriore metodo di valutazione dei modelli è stata la 10-fold cross validation, che permette di calcolare l'accuray media per un determinato modello di machine learning. L'implementazione di questo metodo ha dovuto tenere conto del fatto che le istanze del dataset in questione sono distribuite lungo un anno, quindi una 10-fold cross validation senza cambiare l'ordine dei valori avrebbe creato una situazione molto vicina alla divisione del dataset in mesi. Questo avrebbe portato ad avere modelli trainati su determinati mesi ma testati su altri. In questi casi il training set non sarebbe stato rappresentativo dell'intero database e, quindi, questo avrebbe condizionato i risultati. La soluzione adottata è stata quella di randomizzare l'ordine delle istanze del dataset prima di suddividerlo in 10 subset.

Facendo sempre riferimento alle tabelle sopra riportate, si può notare come l'accuracy media dell'albero di decisione, calcolata attraverso la 10-fold cross validation, risulti più alta rispetto a quella della rete neurale.

Però, osservando attentamente, si può notare che il range di valori compreso tra Min_Acc e Max_Acc sia molto più elevato nell'albero di decisione rispetto alla rete neurale: questo è dovuto al fatto che, al variare della proporzione tra istanze con label "in orario" e "in ritardo", la classificazione risulterebbe più o meno difficile. Infatti, aull'aumentare del numero di istanze realmente "in orario", la classificazione otteneva risultati migliori poiché è più semplice classificare istanze di questo genere. Mentre nel caso opposto si ottenevano risultati di accuracy peggiori a causa della maggior difficoltà nella classificazione delle istanze "in ritardo". Questo non accade all'interno della rete neurale, ciò comporta una maggior robustezza del modello. Infine, considerando l'accuracy dei modelli riportati (Model_Acc), la rete neurale ottiene risultati migliori.

10 CONCLUSIONI

10.1 Data Technology

Per quanto riguarda questa parte del progetto, possiamo ritenerci molto soddisfatti del lavoro svolto. I dataset iniziali erano tutt'altro che pronti per essere utilizzati: dopo una buona fase di analisi che, tra le altre cose, ci ha permesso di capire come indirizzare tutto il lavoro futuro, abbiamo riscontrato un numero così elevato di problemi di qualità da rendere uno dei due dataset praticamente inutilizzabile. Grazie ad un'attenta analisi di svariate dimensioni di qualità, siamo infine riusciti a colmare tutte le lacune presenti inizialmente nei dataset. Tutti i tipi di errori sono stati descritti, valutati e corretti utilizzando il linguaggio di programmazione R; tutto il codice generato e utilizzato per la pulizia e il trattamento dei dati è consultabile e pronto all'utilizzo nella cartella "Data_Technology\Code". Inoltre, tutti i procedimenti logici e pratici fatti durante lo svolgersi del lavoro sono riportati nell'apposita sezione di questa relazione e riassunti nei commenti del codice e nei ReadMe presenti in ogni sottodirectory del workspace.

Grazie alla documentazione fornita è quindi possibile replicare il processo che ci ha portati ad ottenere un dataset finale, corretto secondo tutte le dimensioni di qualità considerate, a partire dai dati raw inizialmente raccolti.

10.2 Machine Learning

La parte del lavoro che riguarda la classificazione delle istanze del dataset, invece, ha ottenuto risultati migliorabili. Il lavoro svolto, a partire dalla scelta del sottoinsieme di istanze da utilizzare per la fase di machine leaning, è disponibile per il testing, documentato e adeguatamente commentato. È possibile seguire il workflow grazie alla documentazione che ne descrive i dettagli, sia per quanto riguarda i problemi riscontrati, sia nei ragionamenti che hanno portato ad ottenere le soluzioni. Durante lo sviluppo di questa fase sono state utilizzate due diverse tecniche di machine learning proprio per ottenere risultati diversi tra loro nei valori e nel significato. Lo scopo finale non era quindi solo ottenere i risultati migliori, ma anche metterci in condizione di comprendere le motivazioni di tali risultati.

Il principale punto debole di questa parte, purtroppo, è proprio il risultato finale.
Questo è dovuto al tipo di problema preso in considerazione: sarebbe stato possibile migliorare il risultato finale, ma abbiamo deciso di mantenere la natura di questo problema di classificazione il più reale e utile possibile. In un caso come questo non solo la correlazione tra la classe di appartenenza di un'istanza e i valori dei suoi attributi è molto bassa, ma è realistico pensare che esistano molti altri fattori, non presenti nel dataset e nemmeno prevedibili, che possano influenzare una classificazione di questo tipo.

10.3 Idee per miglioramenti futuri

In conclusione si sono considerate diverse possibilità che potrebbero portare ad ottenere risultati finali migliori. La realizzazione di queste idee non è stata considerata a causa della mole di lavoro e del tempo necessario all'implementazione. Le principali sono:

- Ristrutturare il dataset: lo scopo di questo lavoro porterebbe a modellare la struttura del dataset in modo da rendere possibile ed efficiente l'aggiunta di nuove istanze, anche con un diverso numero di attributi.

- Ottenere nuovi dati: l'aggiunta di nuove istanze e nuovi attributi riguardanti il caso di studio potrebbe portare ad un miglioramento delle performance dei classificatori.

- Utilizzare una maggior potenza computazionale: per motivi tecnici non ci è stato possibile testare determinati modelli di reti neurali e, soprattutto, utilizzare la totalità del dataset costruito durante la parte di data technology.